題名：「意味処理：文解析と文生成」
-----第一論文補稿-------

第１章　まえがき

前の論文「意味処理」では、動画システムの紹介に重むきを置いたので、文解析と文生成の技術にはあまり触れなかった。本論において、柔軟な文解析と文生成はどんなふうに実現できるかを議論したいと思う。

基本的に情報の単位で文を解釈していくという方針を採る。情報の単位については、第２章で説明している。そうして、この手法は、なにごとも人間がしているように計算機でも行おうという、人間の持つ柔軟性に迫る手法の一つであることを訴えたい。

自然言語処理を難しくしているのは、それが持つ曖昧性である。それを第３章で整理している。そして、第４章で曖昧性を克服する文解析を示し、第５章に文生成の手法を提案する。

自然言語処理の手法、技術はさまざまであり、豊潤な技術体系に成長している。本論がその一つの技術として採用されることを願う。

第２章　情報の単位

情報の単位とは、文章を頭から解釈していくときに、意味のまとまりとしてまとめて抽出し記録していくべき、最小の単位である。それは、基本的に、名詞と格助詞（助詞相当語)、動詞と助動詞（助動詞相当語）をまとめたものである。慣用句とか、連結語とかもまとめて、情報の単位とする。

例をあげる。

（１）名詞と格助詞
　「私は」、「花を」など
（２）動詞と助動詞
　「行ってしまった」など
（３）慣用句
　「以下の条件の時」など
（４）連語
　「見たことがある」、「行ってきた」など

その他、形容詞とか副詞の簡単なものは名詞とか動詞に結びつけて、情報の単位とするのが分かりやすい。

以下に、銀河鉄道の夜の一節を取りあげ、情報の単位を説明する。情報の単位を意味解釈の基本単位とすることの利点を納得されたい。

【例文】
「/窓の外の、/まるで花火でいっぱいのような、/天の川のまん中に、/黒い大きな建物が/四棟ばかり/立っていて、/その一つの屋根の上に、/目も覚めるような、/サファイヤとトパースの/大きな二つの透き通った球が、/輪になって/しずかにくるくると回っていました。/」
（説明）情報の単位の切れ目に「/」を入れている。
・「窓の外の」は名詞「窓」と助動詞相当語「の外の」の固まりである。この助動詞相当語は、英語前置詞「out」を思い浮かべれば、納得できるだろう。
・「まるで花火でいっぱいのような」は、副詞「まるで」と名詞「花火」、それに助動詞相当語「でいっぱい」と助詞相当語「のような」の連結で、ひとまとまりの意味を表している。
・「天の川のまん中に」は、名詞「天の川」と助詞相当語「のまん中に」の固まりである。
・「黒い大きな建物が」は、形容詞「黒い」、「大きな」と名詞「建物」および助詞「が」が連結している。
・「四棟ばかり」は、名詞「四棟」と副詞「ばかり」の連結だが、これは数詞表現で、形容詞のように前出の名詞を修飾する。だから、「黒い大きな建物が四棟ばかり」で一つの情報の単位とすべきかもしれない。
・「立っていて」は、連結動詞「立ってい」と助詞「て」の固まりである。
・「その一つの屋根の上に」は、指示詞「その」と数詞「一つの」と名詞「屋根」、助詞相当語「の上に」の固まりである。
・「目も覚めるような」は慣用語である。
・「サファイヤとトパーズの」は名詞「サファイヤ」と「トパーズ」の並置に助詞「の」がつき、形容詞相当語になっているものである。
・「大きな二つの透き通った球が」は、形容詞「大きな」と数詞「二つの」、形容動詞の連体形「透き通った」に、名詞「球」が続き、格助詞「が」でこの成分の働きを表しているものである。
・「輪になって」は慣用表現である。
・「しずかにくるくると回っていました」は、副詞「しずかに」と「くるくると」と連結語「回っていました」が連結したものである。

このように、情報の単位は、独立する意味と、文中での独立した一つの機能を明快に表現する単位となっている。人は、この単位を戦略ポイントと位置づけて、自然言語処理をしているのだろう。日々の言語活動を反省してみると、そんな気がしてくる。

本論文では、情報の単位を梃子にして、文解析、文生成をしていく方法を考察していく。

情報の単位の部分では、曖昧性解析必要性は少ないことに目をつけたものである。

第３章　自然言語の曖昧性

情報の単位がいつも唯一切り出せ、意味も機能も一つなら自然言語処理はらくなものである。そうはいかない。現実問題、情報の単位の決定にも、情報の単位の係り受け関係にも曖昧性がある。また、欠落した情報というものもある。主語が無い、目的語がない、それらを文をまたがって調達してこなくてはならない。そんな問題もある。これらの問題を解決してこそ、完全な自然言語処理と言える。

３．１　曖昧性議論

曖昧性にはつぎの３つがある。

（１）係り受けが曖昧である

「美しい庭の石」と言ったとき、「美しい」は「庭」を修飾するのだろうか、「石」を修飾するのだろうか。曖昧である。

（２）格支配が曖昧である。

「象は鼻が長い」といったとき、主格は「象」か、「鼻」か。曖昧である。

（３）情報が欠落している。

・「それは彼がくれた」といったとき、主語「私」を補わねばいけない。

・「私はウナギだ」といったとき、意味がとれない。

a.「何を食べる？」の答えならば、「ウナギを食べる」といっているのだし、

b.「君は誰だ？」の答えならば、「ウナギという魚だ」ということになる。

日本語では、主格、目的格がよく省略されるので、特に注意して対策を講じなくてはいけない。

日本語、英語もそうであるが、先頭から読んでいって分かるように、文はできている。係り受けも、省略単語も容易に想像できるように、ある程度の規則性を持ち合わせているものなのである。悪文は例外として、こなれた文は綺麗である。

（例文１）私は、彼から貰ったリンゴを学校へ持っていった。

という文は、次の括弧で示すような入れ子構造を取っている。それも左から右に流れる形をしている。これで、左から意味が確定するので、すっとその意味がとれるのである。

格助詞「は」は主題（主語）の提示を示す。「・・・へは」とか 「・・・をは」とか、関連拡張表現がある。

（例文１の入れ子構造）

（主題：私は）（（彼から貰った）リンゴを学校へ持っていった）。

３．２　係り受けの曖昧性

形容詞の係り受けは、次の文ではなんともいえない。

（例文２）美しい薔薇の花壇

形容詞「美しい」が「薔薇」に係るのか、「花壇」に係るのか不明である。どちらもありそうな表現だからだ。

次の文では、明確である。
（例文３）美しい西の丘
形容詞「美しい」は「丘」に係る。「西」には係らない。それは「美しい西」は非文であるからである。

単文についても同じような事が言える。
（例文４）空を飛んでる船からみたツバメのようす
単文「空を飛んでる」は明確にツバメに係ることがわかる。それは「船が空を飛ぶ」は非文で、「ツバメが空を飛ぶ」はありそうなことであるからだ。

このように、文法を越えて、意味世界に持っていって、始めて係り受けが解決できることは多々あることである。文解析に知識が必要な所以である。
知識としては非文か非文でないかの判定をできるものということになる。それには、「船」として、なにか「movie in warter」なんてプリミティブ属性を持たせて、それで、空を飛ばないという推論を働かせるというような議論ができる。しかし、そんな知識は無数にあるので、手作業で構築するのは困難である。システムが自動学習機能を持ってくれれば、そのようなプリミティブ議論は現実味がでてくるが、現在学習システムは実用になっていない。現在でできそうなことは、コーパスを収集して、そこになければ非文と判定する、というようなことである。コーパスも広い意味で自動学習機能の言えるから、それはそれで、学習機能を持った文解析システムといえるかもしれない。

また、情報の欠落もある程度、前後の文から埋め合わせることができるものである。
（例文５）ロロは私の飼い犬で、一緒によく散歩したものだ。しかし、鎖につないでの散歩は、つかれるものだった。もうやたらと匂いを嗅ぎまくって、ぐいぐい道の端にひっぱっていって、動かなくなってしまうのだ。

さて、例文５の各文の主語は何でしょうか。「つかれるものだった」の主語は「私」、「動かなくなってしまう」の主語は「ロロ」。
この判断は、「つかれるものだった」という言い回しから、主語が「私」と判断されるのだ。「つかれたようだった」なら、主語は「ロロ」と判断される。この文の登場人物は「ロロ」と「私」であることから、この２者のどちらがらしいかの判断となる。基本的には、格助詞「は」で、「ロロ」にフォーカスがあることを提示されているのであるから、主語としては「ロロ」になることになる。「動かなくなってしまう」の主語は、そうして、「ロロ」と判断するのである。

曖昧性は、あらゆる手がかりを駆使して解決していくものである。これだけの情報、知識があれば十分というものはない。システムが、どのような情報、知識を利用できるかという事柄は、システムの決定的な能力の指示子ということだ。将来ロボットができて、その能力を議論するとき、こうしたことが基準となって定量的に評価されるようになろう。
そのとき、知識の表現を規格かし、知識を整理しておくことは、知識の共有のために重

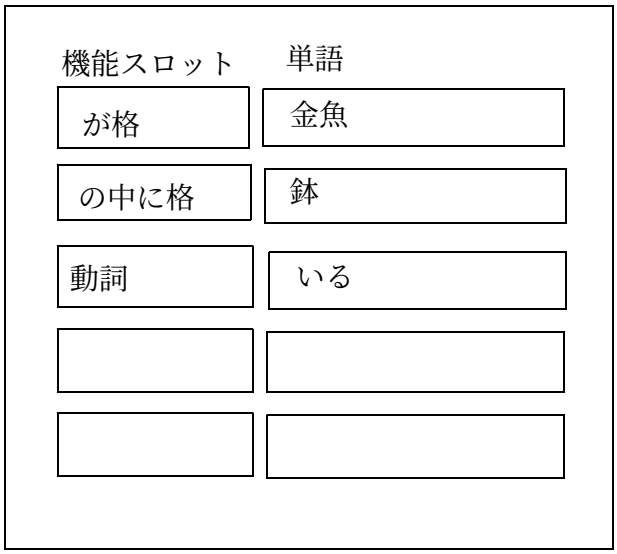
要となる。知識は記号で定義し、表されるはずだ。プリミティブな記号セットを定義し、表現は擬似自然言語で曖昧性無く、パターンマッチングで同値性を確認できるような厳密なものとする。そうしたことが基本だろう。

係り受け解析で、前後の単語との絡みで、判断していくべきものがある。
（例文６）美しくて赤い金魚
（例文７）目玉が美しくて赤い金魚
例文６では、形容詞「美しく」は名詞「金魚」に係る。例文７では、「美しく」は「目玉」に係る。この判断は、「美しく」の先行詞「目玉が」があるかないかという状況判断である。

以上をまとめると、意味処理としては、次の４点が大切であると結論できる。
（１）非文チェック
（２）登場人物の把握と、フォーカスはなにかの把握
（３）先行詞、後続詞の把握
（４）登場人物の能力、行動の性向など、付加情報の把握

第４章　文解析

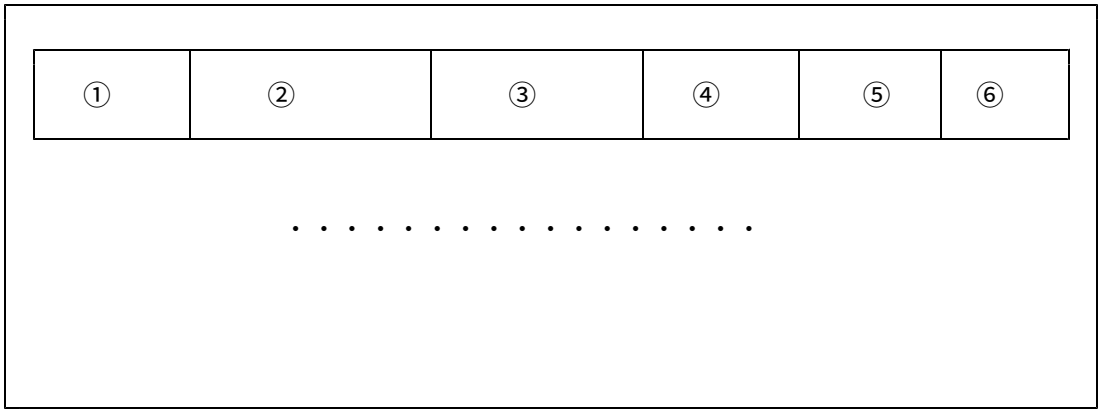
文解析を情報の単位に重点を置いて実施するとしたときの方法論を示したい。
フレームとそのデータ要素であるスロットというものを考えると処理のイメージを考えやすい。受け付ける自然言語は単語や機能語などの要素に分かれ、機能に応じて、特定の位置に納まっていくからである。フレームを自然言語処理に用いるのは素直なことなのである。ただ、自然言語の要素は数が不定なので、実戦にはリスト構造を用いることになる。



「金魚が鉢の中にいる」のフレーム

図４．１　フレームとその中のスロット

文を表すフレームには、基本的に、単語（名詞、動詞、形容詞、副詞）と機能語（助詞、助動詞、動詞、相当語）が対でスロットに納めていくことになるが、作業の都合上、修飾先を示す情報と先行詞情報、後続詞情報をつける。意味処理も行うなら、そのためのスロットも用意することになる。



①:機能スロット
②:単語スロット
③:修飾先情報スロット
④:先行詞情報スロット
⑤:後続詞情報スロット
⑥:意味情報スロット

図４．２　実戦的なフレームの構造

形容詞、副詞には修飾先（掛かり先）というものがある。形容詞は名詞を修飾するし、副詞は動詞や形容詞、副詞を修飾する。

動詞には持ち得る格というものがある。格は動詞によって決まってくる。

・「私は学校へ行く」の動詞「いく」はへ格を取る。

・「私は君に本をあげる」の動詞「あげる」はに格とを格を取る。

その他に文の中に孤立して、時間格、場所格というものが現れる。これは動詞には依存しない。シーンの設定されている時間と場所を表すのである。

場合に寄っては、フォーカス情報も持つ方がよい。格助詞「は」によって、主題が提示されていることを処理として利用したいことは多々あるからである。なお、登場人物の一部だけが主題になるとは限らないことに注意。「私がやったことは」というように、文が主題になることがある。これを考慮しておくことが大切だ。

こうしたことを知ると、動詞には格情報を、形容詞、副詞には修飾先情報を持たせておく必要があることに気づく。そうした情報もフレームとしてもち、これをテンプレートと呼ぼう。

テンプレートを見ながら入力文を解析して、文フレームの各スロットを埋め込んでいく

のである。

手順はこうである。例文７で処理のアウトラインを掴んで欲しい。
（例文７）私は、長野駅で買ったジュースを飲みながら、新幹線の旅を楽しんだ。

用意するものは、総合的解析結果を入れるフレームと、作業用に構成した単文を入れるフレーム３枚である。
（１）「私は」を読む。は格として、「私」を総合解析フレームの先頭のスロットにいれる。意味情報に「主題」と明示する。
（２）「長野駅で」を読む。第一作業フレームの先頭スロットに、で格として「長野」をいれる。同時に総合解析フレームに、「私は」の情報に続いて、で格として「長野」を入れる。意味情報にレベル１（第一単文の要素であることを示す）を付す。先行詞は、が格とする。
（３）「買った」を読む。また、次の語「ジュース」を先読みする。動詞「買う」のテンプレートを見て、を格を取ることを知り、「ジュース」が第一作業フレームの、を格であることを推論する。また、非文チェックを入れる。非文でないことが確認される。第一作業フレームに、を格として「ジュース」を入れる。そして、第一作業フレームは動詞を読み込んだことで、閉鎖する。
総合解析フレームに、動詞として「買った」を入れ、これの意味情報にレベル１を設定、修飾先を「ジュース」とする。「長野駅で」の修飾先には動詞「買った」を入れる。意味情報にレベル１を付す。
（４）「ジュースを」を読む。第二作業フレームの先頭スロットに、を格として「ジュース」をいれる。同時に総合解析フレームに、「買った」の情報に続いて、を格として「ジュース」を入れる。先行詞は動詞であり、単文であるとも設定する。意味情報にレベル２を付す。
（５）「飲みながら」を読む。動詞「飲む」のテンプレートを見て、を格を取ることを知り、第二スロットが主格以外は完備となっていることを確認する。非文でないことも確認する。第二作業フレームは動詞を読み込んだことで、閉鎖する。
総合解析フレームに、動詞であり、格として「ながら」も取ることを示し、「飲む」を設定する。「ジュースを」のスロットの修飾先に「飲みながら」を設定する。意味情報にレベル２を付す。
（６）「新幹線の」を読む。助詞「の」の存在から、名詞に係る修飾語となることを、意味に付加し、総合解析フレームに「飲みながら」に続けて、入れる。先行詞に動詞と単文を設定する。意味情報にレベル３を付す。
（７）「旅を」を読む。第三作業フレームの先頭スロットに、を格として「旅」を設定する。同時に、総合解析フレームに、「新幹線の」の情報に続いて、を格として「旅」を入れる。「新幹線の」の修飾先として、「旅を」を設定する。「新幹線の旅」の非文チェックも入れる。意味情報にレベル３を付す。
（８）「楽しんだ」を読む。第三作業フレームに動詞「楽しんだ」を入れる。を格を取っていることを知り、第三スロットが主格以外は完備あることを確認する。非文でないこと

も確認する。第三作業フレームは動詞を読み込んだことで閉鎖する。
　総合フレームに、動詞であり、文の末尾であることを示して、「楽しんだ」を設定する。
「飲みながら」の修飾先に「楽しんだ」を設定する。意味情報にレベル３を付す。
「旅を」の修飾先に「楽しんだ」を設定する。

　以上で、文解析は終了する。これで、文の要素と各要素間の係り受け関係が明確になっている。
　この例では、非文があったりして生じる曖昧性が出てこなかったが、曖昧性を発見したら、フラグを立て、バックトラッキングをして、再度、フレームを構成しなおすことになる。また、単語不備の場合は、その旨を総合解析フレームにナルスロットとして明示しておき、後段の処理に任せることも行う。

第５章　文生成

　文生成には、プリミティブな情報だけでは不足する。付加情報が重要となる。move というプリミティブだけでなく、walk とか、run といった付加情報で、「歩く」とか「走る」とかの動詞表現を決定できるのだ。
　生成には文体もあるし、動詞の格支配という文法もある。文の形を示すテンプレートが必要である。
　手順としては、先ず、形容詞、副詞の修飾成分の解決を行い、単文テンプレートで文要素を生成する。それは、動詞起動型の処理になる。次に、できた単文の修飾関係情報を用いて、一文に合成する。追加の単語を生成したり、重複して不要になる（単文が名詞を修飾する場合）ものを削除する。いわゆる文の変形作業を行うことになる。
　文生成は文解析ほど難しくない。曖昧性が無いからである。

第６章　まとめ

　文解析と文生成について考察してきた。自然言語のキーポイントといえる技術であるから、多くの研究があるし、実働しているシステムも多い。本論では、情報の単位を基本的処理の要と位置づけ、柔軟性のある方法を探った。そこではフレーム技術が重要となってくる様子が見えてきたと思う。
　これによって、第一論文の足りなかった議論の一部の追加を終えることとする。