自然言語処理 101−18
(1994. 5. 28)

意味情報の主導による不適格文の処理

永松健司 田中英彦

東京大学 工学部

実際の文使用の状況においては、構文的・意味的な誤りを含んだ文章や、人間から見ると正しくても自然言語処理システムの解析能力を越えた文章が入力される場合がある。人間なら、そのような不適格性を含むような文章でも、個々の部分の意味を考えることで、その文全体の意味を理解することが可能である。本研究では、部分的な文の意味の整合性をチェックすることで、不適格性を含むような文でも、その全体の意味を解析できるシステムを試作した。

Processing of Ill-formed Sentence based on Semantic Information

Kenji NAGAMATSU, Hidehiko TANAKA

University of Tokyo
7-3-1 Hongo, Bunkyo-Ku, Tokyo, 113, Japan

In the practical language use, the sentences are used which contain some syntactic or sematic ill-formnesses, or which an NLP system can't process. We, human beings can recognize those ill-formed sentences by considering the meaning of the parts of them. In this paper, we will discuss the approach of our system that can process ill-formed sentences by checking the semantic informations of partial sentences.

1 はじめに

自然言語インタフェースシステムにおいて、直接、人とのやり取りを行なう場合、ユーザはいつも構文的・意味的に正しく、かつ、そのシステムが解析することの可能な文のみを入力するとは限らない。それよりもむしろ、冗長な挿入句や、語句の脱落などを含む非文法的な文章(不適格文)はかなり高い確率で現れると考えられる。これらの文章に対して、単に棄却してしまうだけではユーザに不親切というだけでなく、ユーザの発話した文の情報および、文の途中までの解析に要した処理を無駄にしてしまうことにもつながるであろう。

これに対して、我々、人間ならば、ある程度、構文的・意味的な誤りを含んだ不適格文であっても、その断片的な句などの意味から、全体の文の意味を理解することができる。このことから、不適格文の解析では、意味情報を最大限に活用して解析にあたることが有効であると思われ、部分的な文の意味を基に全体の文の意味を推定するという形で、不適格文でも棄却することなく解析できるようにした自然言語解析システムのプロトタイプを試作した。本稿では、この解析システムの基本的アプローチについて述べるとともに、この処理システムで行なわれる処理の概要について説明する。

2 不適格文解析の基本的手法

この章では、今回試作した処理システムで採った基本的なアプローチ、すなわち処理のモデルについて述べる。また、不適格文自体の処理の考え方についても説明する。

2.1 解析処理の基本的アプローチ

前章で述べたように、文の中の不適格性に出会うと、通常、我々人間は、その前後の語句の、または全体の文の、以前の文の意味と照らし合わせて、意味的に整合性のある解釈を生成して、不適格性を回避していると思われる。

例えば、

例1*) 誤って挿入さいれた文字がある。

という文では、人間なら「挿入された」であると解釈できるが、通常のシステムでは、「さいれた」の部分が構文的に解析不可能となってしまう。

これは、人間はあらかじめ意味を先読みしており、構文的・意味的に予想と反する部分(すなわち不適格性を有する部分)が現れた場合に、文全体の意味が整合性良くなるような解釈を取るようにしているためと推測される。

この考えを基に、図1のような、文の意味を先読みする処理モデルを設定した。ここでは、入力された文に対して、まずその意味の概要が解析され、その概要を参考にして通常の文解析(図中では詳細な解析)が行なわれる。概要把握の段階では、その文で主張されている意味の概要を抽出するだけで、構文的な解析も含めて細かい部分の解析は行なわれない。また、次にどのような文が現れるかについて予測を行ない、その情報も解析に役立てる。

2.2 不適格文処理の手法

前節のような枠組によって、解析の途中で不適格性を検出した時点で意味情報を利用することができるようになる。では、実際に不適格性にどのように対処したらよいかについて考えてみる。

例1で見たように、人間はあらかじめ文章の意味の概要を理解した後で不適格性に対処していると思われる。その対処の仕方としては、

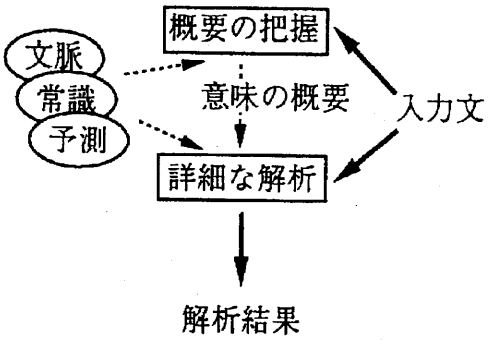

図 1: 意味を先読みする処理モデル

1. 不適格性の種類を判断して、その誤りを回復した文章に変換してから理解する。

2. 文章の言わんとしている意味になるような文を生成して、実際の文章との比較を行なう。

3. 不適格性を含む部分の理解をあきらめ、先に把握した概要のままの意味とする。

などが考えられるだろう。また、このうち 2. は、1. の不適格性の種類判定の前段階として行なわれるかもしれない。

今回、試作した処理システムでは、上のうちの 1. と 3. に対応する処理を行なうようにしている。不適格性の種類判定の方法として、2. のような処理を行なうことが必要な場合もあるかもしれないが、今回はプロトタイプということで見送った。今後、必要ならば検討してみたい。

1. によって、種類の特定できる不適格性については、誤りを訂正して正しく解析することが可能である。また、種類の特定できない不適格性についても、3. により、先に文脈や予測によって解析した意味の概要を用いて、文全体の尤もらしい意味を解析することができるようになる。

3 解析システムとその処理方法

この章では、試作したシステムの構成、および各部でどのような処理が行なわれるかについて述べる。

3.1 システムの概要、およびその構成

システムの処理のおおまかな流れを図 2 に挙げる。入力された文はまず、概要の意味が解析され、その後、その意味を利用して、詳細な解析が行なわれる。その時に不適格性が検出されなければ、その解析結果が最終的に出力される。しかし、途中で不適格性が検出された場合は、その不適格性の特定を行ない、特定されればその回復処理を行なう。不適格性が特定できない場合は、最初に概要の意味として解析されたものがその部分の解析結果として出力される。

このシステムでは文の意味情報として、格フレームによる記述を行ない、解析結果もその形で出力する。また、ある文の次にどのような文がくるかを、図 4のように、その格フレームのネットワークの形で記述しており、対話モデルとして次文の予測、および文脈間での名詞の照応関係の取得に利用している。現在、この対話モデルは人間が手で抽出して与えたものであり、書店における客と店員の会話 (の内の客側の発話) をモデル化している。

処理システムのブロック構成は図 3のようになり、Prolog 上に実装されている。また、形態素・構文解析における辞書としは、EDR の日本語単語辞書を利用している。

3.2 概要の意味の解析

入力された文はまず、その意味の概要が解析される。これは文の大まかな意味を先読みする目的で行なわれ、構文解析は行なわずに、キーとなる名詞および動詞を抽出

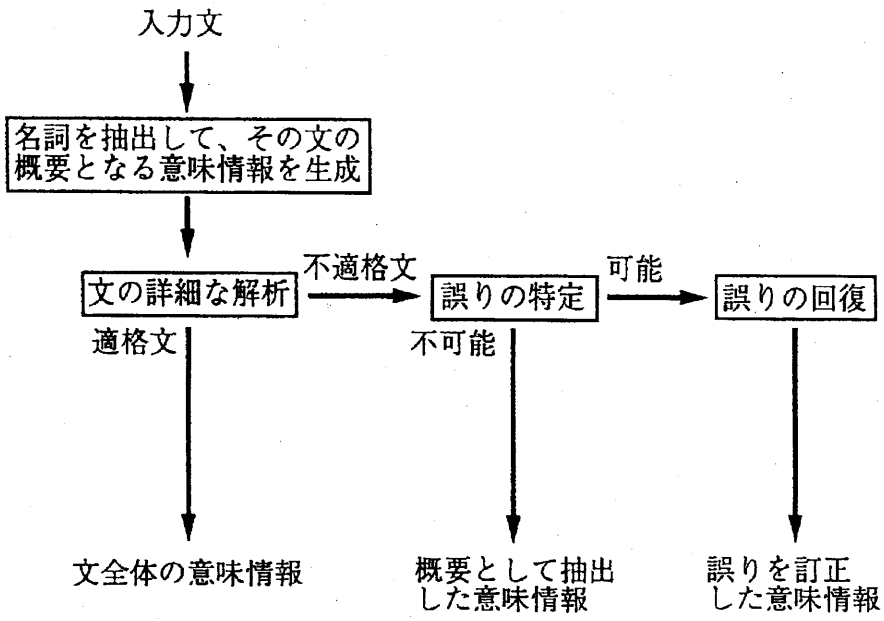

図 2: 解析処理の流れ

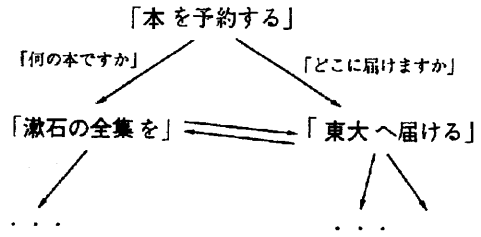

図 4: 対話モデルの概念図

して、それを基に文全体の大まかな意味情報を生成する。その方法は次のようなものである。

1. 文の先頭から、名詞または動詞となるものをすべて抽出する。

2. 対話モデルで次に現れると予測される文の中から、1. で抽出した名詞・動詞ともっともよくマッチングするものを選択する。そのマッチングの基準としては、名詞よりも動詞が一致するものを高い得点として、もっとも一致の度合の大きいものを選択する。

   その場合、マッチングの優先順位は、

   動詞の一致 > 類語の一致
   > 意味素性の一致

   の順である。

3. 選択された文の意味に対して、対応する名詞・動詞を 1. で抽出したものに置き換える。

このようにして、入力された文に対して、その概要に対する意味情報が生成される。

3.3 取り扱う不適格性の種類

前述のように、いくつかの不適格性については、その種類を特定し、適切な誤り回復処理を行なう必要がある。文章中に現れる誤り、つまり不適格性と言っても、その種類は多種多様にわたる。このシステムでは、それらのうちから代表的なものを取り

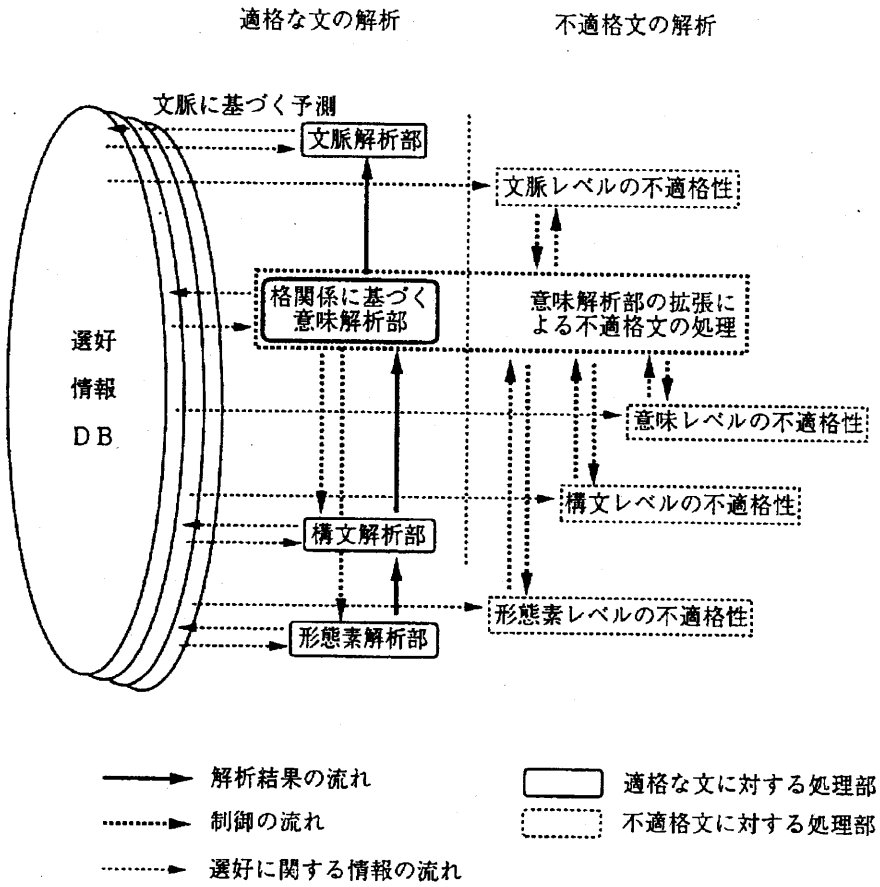

図 3: システムの構成

上げて、それらを含む不適格文についてのみ取り扱うことにする。

まず、大きく、形態素・構文レベルにおける誤りと意味レベルにおける誤りに分けられる。そして、今回のシステムで取り扱う不適格性の種類としては、それぞれ次のようなものを取り上げた。

- 形態素・構文レベルでの不適格性

挿入誤り

例 2*) 誤って挿入さいれた字がある。

(判定条件) ある文字・語句の前後で接続が不可能、かつ、その文字・語句を取り除くと前後の接続が可能となる場合

(回復処理) 挿入された文字・語句を取り除く。

脱落誤り

例 3*) 間違って落され文字がある。

(判定条件) 前後の接続ができなくなる位置があった場合

(回復処理) 構文的にそこに補う語句が特定できれば、それを補う。名詞や動詞など品詞だけしか推定できない場合は、概要の意味情報から、位置的に対応する名詞・動詞があれば、それで補う。特定できない場合は、意味情報中でそれに対応する部分は省略される。

置換誤り

例4*) ある文字が別が文字で置き換えられた。

(**判定条件**) ある文字・語句の前後で接続が不可能、かつ、その文字・語句を取り除いても前後の接続は不可能である場合

(**回復処理**) その文字・語句を取り除いた後、脱落誤りと同じ回復処理を行なう。

- 意味レベルでの不適格性

必須格の欠落

例5*) 抜けている。

(**判定条件**) 必須格である語句が省略されている場合。これは文全体の解析が終った時点で必須格が揃ったかどうかのチエックを行ない判定する。

(**回復処理**) 概要の意味情報から、対応する格情報が補われる。

意味制約違反

例6*) 意味が飛んでいる。

(**判定条件**) 名詞句の解析が行なわれた時点で、名詞—動詞間の意味制約のチエックを行なう。

(**回復処理**) 名詞句を動詞に優先して残すようにし、取り除いた動詞の代わりに概要の意味情報中の動詞でそれを補う。

以上のような種類の不適格性に対しては、このシステムではその特定を行ない、またそれぞれに対して誤りの適切な回復処理が行なわれるようになっている。

3.4 不適格文の処理

概要の意味情報を解析した後には、その意味情報を利用して詳細な解析が行なわれる。ここでは、チャートパーザを用いた構文解析を行ない、名詞句ごとに意味情報を生成して、概要の意味として求められた意味情報の対応する部分を新しく置き換えるという処理を行なっている。

構文解析において解析に失敗した場合、その文は不適格文であると見なされる。不適格文であると判定されると、前節で述べたようなその不適格性を特定する処理が行なわれ、また、特定できたものに対してはその誤りの回復処理が行なわれる。

4 不適格文の解析例

この章では、簡単な例ではあるが、実際の不適格文の処理の例を挙げて、システムの処理の流れを説明する。

今、

例7) 漱石の全集を予約します。

という文が解析された時点であるとし、次に

例8*) 東大の田中研にい届けて欲しい。

という文章が入力されたとする。この文は「い」が誤挿入され、かつ必須格である対象格が省略されている例である。

4.1 予測文の検索

まず、対話モデルから次に現れると予測される文が表1のように検索される。この表で、括弧内の太字の語はその意味に相当する類語が、斜体の英小文字は名詞句が対応することを示す。コロンに続けてその名詞句の意味素性を示す場合もある。また、これらは実際には格フレームの形で記述されているものである。

4.2 概要の意味情報の抽出

次に、入力された文章の概要となる意味情報を抽出する。入力文中から次のような

(x) の (**TITLE**) は zだ
(x) の (**PUBLISHER**) は zだ
(**YOU**) が (x) を (y:location) へ届ける
⋮

表 1:「x を予約する」の次に予想される文

名詞・動詞が抜き出され、先の予測文リスト中の文と比較される。

語句	品詞	意味素性
東大	名詞	location,building,···
田中	名詞	name, human, ···
届ける	動詞	
欲しい	動詞	

ここでは、「届ける」という動詞および、「東大」の意味素性の一致により、

「(**YOU**) が (x) を (y:location) へ届ける」

という文が意味概要情報のテンプレートとして選択される。

最後に、選択された意味情報のテンプレートに対応する名詞句、および過去の文からの対応する名詞句が代入されて、

「(**YOU**) が (**激石の全集**) を
(東大:location) へ届ける」
という意味情報が、この文の意味の概要として生成される。

4.3 不適格性に対する処理

次に構文解析および最終的な意味情報の生成が行なわれるが、構文解析において、

「東大の田中研に / い届けて欲しい。」

ここで解析に失敗する。この場合、「に」と「い」が連接不可能であるためである。よって、この文は不適格文であると見なされ、不適格性の特定処理が行なわれる。

この文の場合、助詞「に」と「い」、「い」と動詞「届け」はどちらも連接不可能であり、かつ「い」を取り除くことで、助詞「に」と動詞「届け」は接続することが可能となる。これによって、挿入誤りであると特定され、挿入文字「い」が取り除かれて、再度、構文解析部へと処理が渡される。

最終的には、
「(**YOU**) が (**激石の全集**) を東大の田中研へ届ける。」
という文に対応する意味情報が生成されることになる。

5 おわりに

本稿では、今回試作した、文の意味情報を基に不適格文を解析する処理システムについて、その処理のアプローチと実際の処理内容を述べた。試作システムということで、細かな処理自体はかなり大雑把 (ad hoc) なものとなってしまったことは否めない。今後、このような細部に関してつめると共に、より一般的な枠組の下で統一的な処理のできる手法を目指したい。

参考文献

[1] 荒木哲郎, 池原悟, 塚原信幸. べた書き日本語文の脱落・挿入誤りの検出法. 情報処理学会自然言語処理研究報告, 1993. 94-7.

[2] 大場健司, 元吉文男, 井佐原均, 横山晶一, 石崎俊. 未定義語を含む文の多段階解析法. 情報処理学会自然言語処理研究報告, 1989. 70-4.

[3] 加藤恒昭. 非文の解析 – チャートに基づく新たな手法. 情報処理学会自然言語処理研究報告, 1991. 83-10.

[4] 佐川雄二, 大西昇, 杉江昇. 対話文における誤りの自動修復. 情報処理学会自然言語処理研究報告, 1993. 93-10.

[5] 塚田孝則, 小柳和子. 未登録語を含む文の一解析法. 情報処理学会自然言語処理研究報告, 1989. 73-6.

[6] 中村孝, 上原邦昭, 豊田順一. 非文法的な日本語文を取り扱う意味主導型解析メカニズム. 情報処理学会自然言語処理研究報告, 1989. 70-8.

[7] (株) 日本電子化辞書研究所. EDR 電子化辞書仕様説明書, 1993.