構文パターンを用いた中国語文解析

王向莉　　宮崎正弘

新潟大学自然科学研究科

1　はじめに

構文解析における曖昧性は大きく分ければ係り受けによる曖昧性（係り受け曖昧性）と構文要素の境界区切りによる曖昧性（境界曖昧性）との二種に分けることができる [1]。構文解析を行う際には、英語では動詞や形容詞の語尾変化、日本語では用言の活用や格助詞などが構文要素の境界情報となっている。しかし中国語は孤立語であり、このような動詞、形容詞の語尾変化はないため、英語や日本語と比べると、構文要素の境界区切りによる曖昧性が非常に顕著であり、従来の文脈自由文法に基づいたパーザを用いて構文解析を行うと、曖昧性は爆発的に発生する [1]　[2]。

本稿では、文全体を取り扱う構文パターンを用いて、境界曖昧性の解消を図る手法を提案する。意味は表現自体がもっている客観的な関係 (表現に結び付け固定された対象と認識の関係) であるという関係意味論 [3] に基づき、文構造は意味の一部であると考え、中国語文表現を徹底的に分析したうえで、多くの構文パターンを作成した。これらの構文パターンを拡張型のチャートパーザである Schart パーザ [4] に実装し、特殊文型、動詞、形容詞、介詞の構文特徴の情報を与えることより、構文要素の境界区切りによる曖昧性の解消を実現した。また、提案した手法の評価を行い、その有効性を検証した。

2　係り受け曖昧性と境界曖昧性の関係

構文解析における二種の曖昧性（係り受け曖昧性と境界曖昧性）の間には、境界曖昧性の発生とともに、係り受け曖昧性が発生するという関係がある。

[w1、w2、w3、w4 w5] という文字列があるとする。

文法規則によると

構造 a：[[w1 w2] w3 w4 w5]

と

構造 b：[[w1 w2 w3] w4 w5]

は

いずれも文法規則に適合する、すなわち、[w1 w2] も [w1 w2 w3] も一つの構文要素として解析可能であるとすれば、境界曖昧性がある。また、[w1 w2 w3] という構文要素内部は係り受け曖昧性がある（図 1)。

構造 a が正解で、構造 b において、構文要素の境界が間違っていると仮定する。構造 b の場合は、[w1 w2 w3] の内部の係り受けによる曖昧性も共に発生する。

したがって、二種の曖昧性は関係がある。係り受け曖昧性は境界曖昧性の発生と共に発生すると言える。境界曖昧性を解消すれば、係り受け曖昧性の発生をかなり抑制することが期待できる。

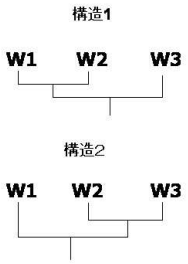

図 1: 構文要素内部の受け係りによる曖昧性

3　中国語文解析における曖昧性の特徴

動詞が語尾変化がなく、しかも文の中で多種の役割を果たすのは中国語の主な特徴である（表 1)。この特徴のため、計算機で構文要素の境界を正しく区切ることが難しく、境界曖昧性が非常に顕著である [1] 。

「自然/n 言語/v 理解/v 研究/v 工作/n 有/v 了/u 新/a 的/u 進展/v」（自然言語理解研究事業に新たな進展があった。）という文を見ると、文脈自由文法による作成した文法規則を使うと、34 もの解析結果を得た。

この文にある 5 個の動詞となりうる語がいずれも述語動詞として解析されてしまった。

表 1: 動詞の役割

動詞	文中の役割	例文
学習	述語	我 学習 日語（私は日本語を勉強する）
学	主語	学 好 語言 非常 難（言語をマスターするのは非常に難しい）
看	目的語	我 喜歓 看 電影（私は映画を見るのが好きだ）
発展	名詞句の中心語	経済 的 発展（経済の発展）
理解	複語名詞	自然 語言 理解（自然言語理解）
学習	介詞目的語	把 学習 日語 （日本語を勉強することを）
起来	方向補語	小王 笑 起来（王さんは笑い始めた）
会	結果補語	我 学 会 遊泳（私は水泳を習って、泳げるようになった）

表 2: 記号と品詞の対応表

品詞の表示	対応品詞
n	名詞
v	動詞
u	助詞
a	形容詞

2で述べたように、境界曖昧性を解消すれば受け係り曖昧性を相当に抑制でき、しかも境界曖昧性が顕著であるのは従来の中国語文解析における主な特徴であるため、境界曖昧性を有効に解消すると全体的に曖昧性を有効に解消できると考えられる。

4 中国語の構文特徴

中国語は屈折語である英語や膠着語である日本語と異なり、用言の語尾変化がないため、品詞と構文要素間に対応的関係が非常に複雑であり、同じ品詞は多種の構文要素を担当するとよく指摘されている [5]。文脈自由文法に基づくパーザで中国語文を解析すると、文法規則による制約がゆるく、曖昧性が爆発的に発生する。本研究では、用言の語尾変化のほか、構文要素の境界情報を提供できるものがあるかどうかを考え、中国語文構造を徹底的分析した。

中国語文は以下の特徴がある。

1. 主語が名詞句、述語が動詞であるのが一般的であり、用言性成分（動詞、形容詞）が体言性構文要素になる場合は、その文構造を制約される。
2. 「是」字文、二重主語文などの特殊文の場合は文構造を制約される。
3. 述語動詞によって、文構造を制約される。
4. 「把」、「被」、「使」、「比」などの介詞を含む文は、その文構造を制約される。

5 構文パターンの作成

4で述べた中国語の構文特徴によって、介詞、動詞、形容詞、特殊文等の構文特徴を与えるため、文全体を取り扱う構文パターンを作成し、境界曖昧性を解消する手法を考えた。ここでは、中国語に最も多い文型である動詞述語文（述語は動詞である文）についてを説明する。

表 3: 述語動詞の分類の一部

述語動詞の記述方式	取れる構文要素	例文
v	主語	車 修 了
v_Cj	主語、補語	車 修 好 了
v_On	主語、名詞目的語	我 学 日語
v_Cf_On	主語、補語、名詞目的語	我 買 回来 一 個 西瓜
v_On_On	主語、名詞目的語、名詞目的語	小王 送 我 一 本 書
v_On_Ov	主語、名詞目的語、動詞目的語	小王 教 我 説 日語

述語動詞によってどのような構文要素（主語、目的語、補語など）をとれるかを考えて、述語動詞を 20 の種類に分類した。表 3 はその分類の一部である。また、

表 4: 構文パターンの一部

述語動詞	構文パターン
v	”(s Sn tp* jf* d* zv* zy* v)”.gram
v_Cj	”(s Sn tp* jf* d* zv* zy* v_Cj Cj)”.gram
v_On	”(s Sn tp* jf* d* zv* zy* v_On uv1* On)”.gram
v_Cf_On	”(s Sn tp* jf* d* zv* zy* v_Cf_On Cf uv1* On)”.gram
v_On_On	”(s Sn tp* jf* d* zv* zy* v_On_On On On)”.gram
v_On_Ov	”(s Sn tp* jf* d* zv* zy* v_On_Ov On Ov)”.gram

表 4 に示すように、それぞれの種類の動詞がとる構文要素の情報をパターンに記述した。

表 5: 記号と構文要素の対応表

s	文
Sn	名詞性主語
tp	時間句
jf	否定判断辞
d	副詞
zv	助動詞
zy	「地字フレーズ」などの状語
uv1	動態助詞 ”了 ”
Cj	結果補語
Cf	方向補語
On	名詞性目的語
Ov	動詞性目的語

6 解析実験結果に対する評価

本構文解析システムの有効性を検証する評価試験を行った。

6.1 機能試験文の作成

作成した構文パターンを Schart パーザに実装し、540 文の機能試験文（表 6,7,8) について机上で検討 した。機能試験文は文法の網羅性を念頭において作成したものである。

表 6: 機能試験文の内容（1）

総単語数	最大単語長	最小単語長	平均単語長
3140	14	1	5.815

表 7: 機能試験文の内容（2）

文の種類	文の数
単文	380
用言が名詞句を係る複文	51
その他の複文	52
重文	57

6.2 実験結果に対する評価

作成した 540 文の機能試験文を用いて、本構文解析システムの評価を行った。540 文のうち、解析可能な文は 537 文、解析不可能な文は 3 文である。図 2 は 540 の試験文について構文解析した際に発生する構文木数がそれぞれ１、2、3、4、5 である文、また構文木が 5 個以上の文の割合を示すものである。図 2 から、構文木が 1、2、3 である文は全部の試験文の 9 割に占め、曖昧性が有効に解消されたことが分かった。また、図 3、4 から構文木数が文の動詞数と単語数の増加につれて増える傾向はほとんどなく、動詞が形態を変えずに文の多種の構文要素を担当することによって発生した境界曖昧性は有効に解消されたことも分かった。

表 8: 機能試験文の内容（3）

文に含まれる動詞数	文の数
0	58
1	112
2	213
3	76
4	34
5	45
6	2

7 まとめ

従来の文脈自由文法に基づいた中国語構文解析において、曖昧性が爆発的に発生する。本稿では、すべての曖昧性を係り受けによる曖昧性（係り受け曖昧性）と構文要素の区切りによる曖昧性（境界曖昧性）とに大別して、この二種の曖昧性の関係を検討し、受け係り曖昧性が構文要素の境界曖昧性とともに発生するこ

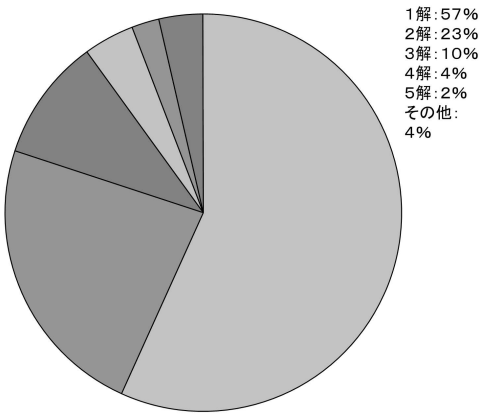

図 2: 構文解析における構文多義発生状況

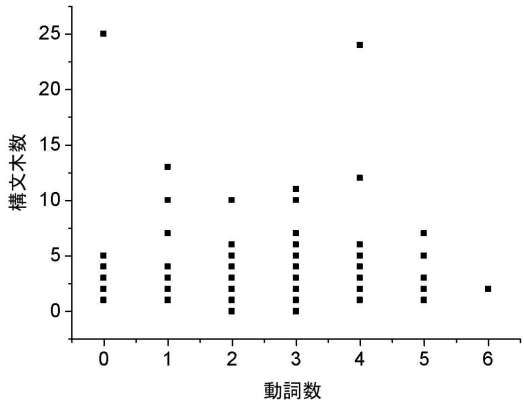

図 3: 曖昧性と文の動詞との関係図

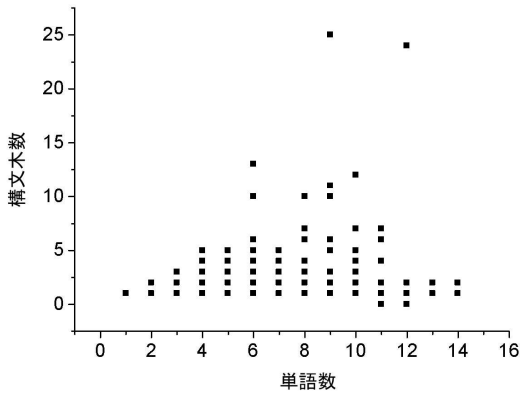

図 4: 曖昧性と文の単語数との関係図

とを述べた。また、中国語は孤立語であるため、英語の動詞や形容詞の語尾変化、日本語の用言の活用や格助詞などの境界情報を与えられなく、構文要素の境界曖昧性が非常に顕著である。そのため、構文要素の境界曖昧性を解消することより、全体の曖昧性を解消する手法を考えた。これは中国語の動詞、形容詞、介詞、特殊文型などの構文特徴情報を全体に取り扱える構文パターンを作成し、これを用いて、曖昧性の解消を図る手法である。

作成した構文パターンを Schart パーザに実装し、本構文解析システムの有効性を検証した。実験結果から、構文要素の境界曖昧性、また、全体の曖昧性は有効に解消されたことが分かった。

本構文解析システムに意味情報や文脈情報を導入し、残された曖昧性の解消を実現することは今後の研究課題である。

参考文献

[1] 王向莉, 宮崎正弘：話者の認識構造を抽出する中国語文パーザ, 電子情報通信学会信越支部大会 I1（2003）

[2] 楊頤明, 堂下修司, 西田豊明：中国語解析システムにおけるヒューリスティックな知識の利用, 情報処理論文誌,Vol.25 No.6,pp.1044-1054（1984）

[3] 三浦つとむ：日本語とはどういう言語か, 講談社学術文庫（1976）

[4] 川辺諭 , 宮崎正弘 ：構造を含む生成規則を扱える拡張型チャートパーザ, 言語処理学会第 11 回年次発表論文集 (2005)

[5] 朱徳熙： 語法答問, 商務印書館 (1985)