中国語意味解析コーパス構築のための句レベルのスコープアノテーション
―文の構成要素の間のコントロール関係の同定および否定の作用域の制御を中心に―

周振†‡　Alastair Butler*　吉本啓*†
*東北大学高度教養教育・学生支援機構　†東北大学国際文化研究科　‡日本学術振興会

要旨
本研究では、ペン通時コーパス式の解析スキームを用いて中国語の実例を解析する。単純な格フレームを超えたより複雑な構文により表現された意味情報を抽出するために、本研究では、文の統語解析の段階において句レベルのスコープアノテーションを行う。これにより、文の構成要素の間のコントロール関係の同定や否定の作用域の制御などに関する複雑な処理が出来るようになった。

1. はじめに

筆者たちは、中国語の無制約のテクストに対して、論理意味表示（述語論理式）を付加した中国語の意味表示コーパスを構築している。その作業は二つの段階に分けられる。すなわち、(1)分析データとして選ばれた中国語の自然テクストに対し統語解析情報を付与すること、および(2)バトラー (Butler 2010) が提唱するスコープ制御理論 (Scope Control Theory; SCT) を実装したシステムで(1)の結果を処理することによる自動的な文の論理意味表示の獲得である。

構造の曖昧性を克服して単純な格フレーム（述語 - 項関係）を超えたより複雑な構文により表現された意味情報の抽出を可能にするために、豊富な統辞情報の提供が出来るアノテーションが求められる。本研究では、中国語の実例により、意味処理の段階で文の構成要素の間のコントロール関係の同定や否定の作用域の制御などに関する複雑な処理をするために、統語解析の段階において句レベルのスコープアノテーションを行う必要があるということを示していきたい。

2. ペン通時コーパス式の解析規約

本研究では、ペン通時コーパス式の解析規約 (Santorini 2010) を使用する。ペン通時コーパス式の解析システムは、ペンツリーバンク式の解析スキームを修正したもので、文の統語構造をラベル付きの括弧で表示する。ラベルには語彙レベルのラベル（N、ADJ など）と句レベルのラベル（NP、ADJP など）の 2 種類がある。文の全ての終端要素（語や助詞など）は語彙レベルのラベルによって、タグ付けされている。一方、句レベルでは形式と機能の両方を指定するタグを付加することがある。例えば、NP-SBJ の場合、NP は句のタイプが名詞句であることを表すと同時に、-SBJ はこの名詞句の機能を主語に限定している。

3. 機能タグを手掛かりとする意味処理

一般的には、依存関係の表示および述語 - 項関係の再構成に必要なため、主語または目的語が動詞の必須格として求められるにもかかわらず文中で表現されていない場合、ゼロ代名詞の追加を行ってそれらを明示する必要がある。

ゼロ代名詞は、一般的に純粋な代名詞類の性質を持つ pro および代名詞類の性質と照応形の性質を合わせ持つ PRO に分類される。従来コントロール構文に対するゼロ代名詞のタギングは、直接 PRO を付け加えることが主流だった。目的語コントロール構文の(1)と主語コントロール構文の(2)に対して、(3)と(4)が示すように、中国語版ペンツリーバンク式の解析スキーム (Xue et al. 2000) では、コントロール補文である IP に対して主語 PRO のアノテーションを行っている。

(1)
张三今晚请我们吃饭。
張三は今晩私たちを御馳走する。

(2)
李四帮我们叫医生。
李四は私たちのためにお医者さんを呼ぶ。

(3)
```
(IP (NP-SBJ (NPR 张三/張三))
    (NP-TMP (N 今晚/今晚))
    (VB 请/招待する)
    (NP-OB1 (PRO 我们/私たち))
    (IP (NP-SBJ *PRO*)
        (VB 吃饭/食事する)))
```

(4)
```
(IP (NP-SBJ (NPR 李四/李四))
    (VB 帮/手伝う)
    (NP-OB1 (PRO 我们/私たち))
    (IP (NP-SBJ *PRO*)
        (VB 叫/呼ぶ)
```

(NP-OB1 (N 医生/お医者さん))))

このようなコントロール構文の解析はある意味で論理的だが、精度の高い述語－項構造をまとめるために、PRO の値を確定しようとする際に問題が生じてしまう。即ち PRO をコントロールしている対象（主文の主語か目的語か）の同定が出来ないのである。

これに対して、本研究では、統語解析の段階で句にそれぞれ異なる機能タグを付与し、IP および CP の下位カテゴリーを精密化することによって、その中に埋め込まれた主語や目的語にデフォルト的解釈を与えることが意味処理の段階で可能となった。

(5)
```
(IP-MAT (NP-SBJ (NPR 张三/張三))
        (NP-TMP (N 今晚/今晩))
        (VB 请/招待する)
        (NP-OB1 (PRO 我们/私たち))
        (IP-INF (VB 吃饭/食事する)))
```

(6)
```
(IP-MAT (NP-SBJ (NPR 李四/李四))
        (VB 帮/手伝う)
        (NP-OB1 (PRO 我们/私たち))
        (IP-PPL (VB 叫/呼ぶ)
                (NP-OB1 (N 医生/お医者さん))))
```

例えば、(5)と(6)では、PRO のタギングが行われず、その代りに、補文の二つの IP はそれぞれ IP-INF（不定詞句）および IP-PPL（分詞句）とされている。SCT により、IP-PPL に対しては、「それによって直接支配される主語項（NP-SBJ 或いは IP-PPL-SBJ）が文中に明示的に表示されていない場合、それより先行ししかもそれと姉妹（それと同じレベルにある要素）の関係を持つ主語項がそれが直接支配する主語項となる」というデフォルト的解釈を（今の段階での暫定的なものとして）与えることが出来る。また、IP-INF に対しては、「それによって直接支配される主語項が文中に明示的に表示されていない場合、それより先行ししかもそれと姉妹の関係を持つ目的語項がそれが直接支配する主語項となる。」というデフォルトの解釈を同様に付与することにより、ほとんどの場合、空範畴を表すゼロ代名詞 PRO のアノテーションを行わずに済む。これによって、文全体の統語構造が簡潔になるとともに、従来のゼロ代名詞 PRO をタグ付けする方法と比べてより精密な要素間の同一指示関係を意味処理によって捉えられるようになる。

(7)
```
∃ x4 t1 e2 e3 (
 x4 = 我们 ∧
  今晚(t1) ∧
   请(e3, 张三, x4, 吃饭(e2, x4)) ∧ tmp(e3) ⊑ t1)
```

(8)
```
∃ x4 x1 e2 e3 (
 x4 = 我们 ∧
  医生(x1) ∧ 帮(e3, 李四, x4, 叫(e2, 李四, x1)))
```

(7), (8)が示すとおり、(1)における主文目的語と補文主語、および(2)の主文主語と補文主語とを同一指示であるとして関係づけられる。

4. 句レベルのスコープアノテーション

さらに、語順変化や否定の作用域などの問題も視野に入れると情況が一層複雑になる。(9), (10)では、主節“张三今天没去学校”と従属節“因为昨晚熬夜了”との順番が逆になっている。今まで紹介してきた解析法を使うと、両方の統語解析結果はそれぞれ(11), (12)になる。

(9)
张三因为昨晚熬夜了今天没去学校。
張三は昨夜徹夜したから今日学校へ行かなかった。

(10)
张三今天没去学校因为昨晚熬夜了。
張三は今日学校へ行かなかった。昨夜徹夜したから。

(11)
```
(IP-MAT (NP-SBJ (NPR 张三/張三))
        (PP (P 因为/から)
            (CP-ADV (IP-SUB (NP-TMP (N 昨晚/昨夜))
                            (VB 熬夜/徹夜する)
                            (AS 了/完了))))
        (CRD *)
        (NP-TMP (N 今天/今日))
        (NEG 没/否定)
        (VB 去/行く)
        (NP-OB1 (N 学校/学校)))
```

(12)
```
(IP-MAT (NP-SBJ (NPR 张三/張三))
        (NP-TMP (N 今天/今日))
        (NEG 没/否定)
        (VB 去/行く)
        (NP-OB1 (N 学校/学校))
        (PP (P 因为/から)
            (CP-ADV (IP-SUB (NP-TMP (N 昨晚/昨夜))
                            (VB 熬夜/徹夜する)
                            (AS 了/完了))))
```

```
        (CRD *))
```

(11), (12)をシステムに入力すると、その意味処理の結果は次の(13), (14)のとおりになる。

(13)
```
¬ ∃ x1 t2 t3 e4 e5 (
      学校(x1) ∧
       昨晩(t2) ∧
        今天(t3) ∧
         因为(熬夜_了(e4, 张三) ∧ tmp(e4) = t2,
          去(e5, 张三, x1) ∧ tmp(e5) = t3))
```

(14)
```
¬ ∃ x1 t2 t3 e4 e5 (
      学校(x1) ∧
       今天(t3) ∧
        昨晚(t2) ∧
         因为(熬夜_了(e4, x1) ∧ tmp(e4) = t2,
          去(e5, 张三, x1) ∧ tmp(e5) = t3))
```

(9), (10)の述語論理式としての(13), (14)には問題点が残っている。(14)では、述語“熬夜”の動作主は x1 の“学校”になってしまっている。それに対して、(13)では、述語“熬夜”の動作主は“张三”であり、これは言語データの実情に合っているが、(13), (14)とも、否定 (negation) のオペレーター ¬ が述語論理式の最初に現れている。そのため、(13)の意味は、「張三は昨夜徹夜したから今日学校へ行ったということは真実ではない」となる。しかし、これは明らかに文の元の意味を捉えそこなっている。

このような望ましくない処理の結果が出た理由としては、まず、CP-ADV（それを直接支配する PP が存在する場合も同様）が直接支配する主語項を、文中にあるそれと同一指示関係を持つ名詞句と同定するために、「CP-ADV によって直接支配される主語項が文中に明示的に表示されていない場合、それより先行ししかもそれと姉妹の関係を持つ目的語項（目的語が存在しない場合は主語項）がそれが直接支配する主語項となる」というデフォルトの解釈を与えたためである。このようなデフォルトの解釈は、ほとんどの情況に対応できるが、(10)のようなその言語が持つ一般的な語順に従わない文を対象とする場合は、矛盾をもたらしてしまう。即ち、(10)では、主文の目的語“学校”が CP-ADV を直接支配している PP“因为昨晚熬夜了”よりも先行ししかもそれと姉妹の関係を持っているにも関わらず、それがその主語にはならず、むしろ主文の主語“张三”こそ補文の主語である。

また、(11), (12)を考察すればわかるように、否定のマーカー“没”はどちらにおいても主文の述語“去”についているため、その述語論理式である(13), (14)では、否定のオペレーター ¬ がみな主節に対して付加されている（その中に従属節が含まれているにも関わらず）。

以上で取り上げた問題点を解決するために、SCT では、第 3 章で見たように句に埋め込まれた主語や目的語に対してデフォルト的解釈を付与する他に、スコープの階層 (scope hierarchy) という概念も導入した。

ペン通時コーパス式の解析規約に従って構築してきたツリーは常にフラットなために、句レベルにおける各要素間のスコープ関係は基本的には自由である。しかし、複数の修飾語・句が存在する場合は、ある程度の制限を加える必要がある。Butler et al. (2013) によると、基本的なデフォルトの順番として、自然言語の語順が参考として用いられる。従って、より最初に現れる修飾語・句はより広いスコープを取ることになる。スコープの階層のデフォルト値に影響を及ぼすプロパティーは、自然言語の語順の他にも幾つかある。まず、文・句の述語（それが補語（CP-THT や IP-INF など）を伴う場合は、その補語）は常にもっとも狭いスコープを取る。また、幾つかの機能タグ（TPC（主題）や VOC（呼格）など）によってマークされる修飾語・句はもっとも広いスコープを持つ。最後に、機能タグ-SBJ を持つ項に対して、他の修飾語・句よりも高いスコープの階層を持つというデフォルトの値を設定する。

以上のようなスコープの階層を決めるためのデフォルト的設定により、自然言語のほとんどの情況に対応できる。更に、(9), (10)のような特別な例外を解決しスコープの階層関係に関する精密な調整を可能にするために、-HIGH および -LOW という二つの機能タグを導入する。以上を踏まえて、スコープの階層を降順で以下のように規定する。

① -TPC、-VOC および-HIGH によってマークされる修飾語・句
② -SBJ を持つ項
③ IP-PPL
④ MD（モダリティー）
⑤ NEG（否定）
⑥ -HIGH および-LOW によってマークされていない修飾語・句（語順順）
⑦ -LOW によってマークされる修飾語・句
⑧ 述語

⑨ 補語

-HIGH あるいは-LOW が付与された修飾語・句が二つ以上ある場合、そのスコープの階層が自然言語の語順で決まることになる。

スコープの階層を導入することによって、CP-ADV も含めて、第 3 章で与えた IP-PPL および IP-INF に関するデフォルトの解釈を以下のように修正・統合することが出来る。

CP-ADV / IP-PPL / IP-INF :
それらによって直接支配される主語項が文中に明示的に表示されていない場合、それらと姉妹の関係を持つ要素からなるスコープの階層において、それらよりも高い階層にありしかもそれらともっとも近い距離を保つ必須項（NP-INST、NP-OB2、NP-LGS、NP-OB1、NP-SBJ）が、それらが直接支配する主語項となる。

以上を踏まえて、(9), (10)の統語解析を以下の(15), (16)のように修正することが出来る。

(15)
```
(IP-MAT (NP-SBJ (NPR 张三/張三))
        (PP (P 因为/から)
            (CP-ADV (IP-SUB (NP-TMP (N 昨晚/昨夜))
                            (VB 熬夜/徹夜する)
                            (AS 了/完了))))
        (CRD *)
        (NP-TMP (N 今天/完了))
        (NEG-LOW 没/否定)
        (VB 去/行く)
        (NP-OB1 (N 学校/学校)))
```

(16)
```
(IP-MAT (NP-SBJ (NPR 张三/張三))
        (NP-TMP (N 今天/今日))
        (NEG-LOW 没/否定)
        (VB 去/行く)
        (NP-OB1-LOW (N 学校/学校))
        (PP (P 因为/から)
            (CP-ADV (IP-SUB (NP-TMP (N 昨晚/昨夜))
                            (VB 熬夜/徹夜する)
                            (AS 了/完了))))
        (CRD *))
```

(15)では、NEG に-LOW を与える。その他は、(11)と全く変わりがない。一方、(16)では、LOW は NEG および NP-OB1 の両方に付与する。これによって、IP-MAT によって直接支配される各要素間のスコープ階層はそれぞれ以下のようになる。

(17)
```
[NP-SBJ (张三)
 [PP (因为昨晚熬夜了)
  [NP-TMP (今天)
   [NP-OB1 (学校)
    [NEG (没)
     [VB (去)]]]]]]
```

(18)
```
[NP-SBJ (张三)
 [NP-TMP (今天)
  [PP (因为昨晚熬夜了)
   [NEG (没)
    [NP-OB1 (学校)
     [VB (去)]]]]]]
```

このように、言語データの実情に合わせてスコープの階層関係を調整することによって、より精度の高い述語論理式が得られるようになる。(9), (10)の意味処理結果はそれぞれ以下の(19), (20)の通りである。

(19)
```
∃ x1 t2 t3 e4 (
 学校(x1) ∧
  昨晚(t2) ∧
   今天(t3) ∧
    因为(熬夜_了(e4, 张三) ∧ tmp(e4) = t2,
     ¬ ∃ e5 (去(e5, 张三, x1) ∧ tmp(e5) = t3)))
```

(20)
```
∃ t1 t2 e3 (
 今天(t2) ∧
  昨晚(t1) ∧
   因为(熬夜_了(e3, 张三) ∧
    tmp(e3) = t1, ¬ ∃ x4 e5 (学校(x4) ∧
     去(e5, 张三, x4) ∧ tmp(e5) = t2)))
```

5. まとめ

本研究ではペン通時コーパス式の解析スキームを用いて中国語の実例をアノテーションした。それだけでなく、意味処理上の要請から考えて、元解析規約では取り扱わないとする部分に関しても、句レベルの要素を対象とするスコープの階層情報の追加など様々な工夫を加えた。これにより、要素間の同一指示関係の同定や否定の作用域の制御など、従来の方法ではなかなか手が届かない領域の意味解析が可能になった。

参考文献

Butler, A. (2010) *The Semantics of Grammatical Dependencies*. Emerald.

Butler, A., et al. (2013) Treebank Annotation for Formal Semantics Research. In Y. Motomura, A. Butler & D. Bekki (eds.), *New Frontiers in Artificial Intelligence*, 25-40. Springer-Verlag.

Santorini, B. (2010) Annotation Manual for the Penn Historical Corpora and the PCEEC (Release 2). Tech. rep., Department of Computer and Information Science, University of Pennsylvania.

Xue, N., et al. (2000) The Bracketing Guidelines for the Penn Chinese Treebank (3.0). Tech. rep., Institute for Research in Cognitive Science, University of Pennsylvania.