中国語統語解析木の形式変換及びその応用に関する研究-Penn Chinese Treebank(3.0)を対象として-

周振‡　Alastair Butler*†　吉本啓†‡
*科学技術振興機構　さきがけ
†東北大学高等教育開発推進センター
‡東北大学大学院国際文化研究科

syusin3@yahoo.co.jp

要旨

自然言語に統語情報を付与したテキストコーパスは、多くの研究分野で大切な役割を果たしているが、その作成は決して簡単な作業ではない。そこで、本研究では、統語コーパスの変換（統語解析木の形式変換）を半自動的に行うための言語の差を越えた一般的な手法を提案する。最初の試みとして中国語を取り上げ、Penn Chinese Treebank (3.0) をPenn Historical Corporaに変換することを考えた。その過程を経て得られた経験の積み重ねとメカニズムの解明によって今後様々な使用目的に対応できる基本データの提供が可能になり、数多くの研究分野に貢献できると思われる。

1　はじめに

自然言語に統語情報を付与したテキストコーパス（以下統語コーパスとする）は、言語学及び情報処理の研究の分析の素材を提供し大切な役割を果たしている。それだけではなく、高精度の形式的意味表示を自動的に得る研究（Butler and Yoshimoto 2012）にとっても欠かせないデータベースになっている。しかし、統語コーパスの構築は膨大なプロジェクトで、決して簡単なことではない。一方、現在のところ統語コーパスの数は何と言っても少なく、作成の目的によって形式上に大きな相違もある。すでに入手可能なリソースを有効的に使用していくためには、統語コーパスの変換を半自動的に行うための言語の差を越えた一般的な手法の確立が重要である。そのために、本研究では、Penn Historical Corporaのスキームをベースに、Penn Chinese Treebank (3.0) を、高精度の形式的意味表示を自動的に得る研究に相応しい形式に変換することを考えている。

2　先行研究

2.1　ツリーバンク（Treebank）

ツリーバンクは、コーパスの一種類であり、データ中の各文に統語情報を付与したテキストコーパスである。代表的なツリーバンクとして、句構造規則に基づいて作成されたPenn Treebankシリーズが挙げられる。Penn Treebank式の解析スキームは英語をはじめ、中国語（Xue and Xia 2000）などに適用されている。しかし、Penn Treebankは元々構文解析器の評価及び訓練を目的としてデザインされたもので、統語木の階層が多く、同じカテゴリーの句や節が何度も重なって部分木を構築することがある。Penn Treebankを用いて形式意味表示を得るためのテキストに対する深い処理を行う際に、述語と述語のとる項との関係の確定は出来るが、節と節の関係（埋め込みか等位か）を決めることは困難である。

2.2　Penn Historical Corpora のスキーム

Penn Historical Corpora 式の解析システム（Santorini 2010）は高い汎用性を持ち、言語ごとに必要なある程度の修正を施した上で、多くの言語に適用されてきた。

Penn Historical Corpora 式の解析システムは、Penn Treebank式の解析スキームを修正したものである。(1)は中国語の例文で、(2)はその統語解析木である。(2)に示すように、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。文のすべての終端要素（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、句レベルのラベルは形式と機能を両方指すことができる。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がこの

名詞句の機能を主語に限定している。

(1)
```
如果 没有 基础 研究 就   没有 科学
もし ない 基礎 研究 直に ない 科学
的       发展。
接続助詞 発展
```
基礎研究なしでは科学は発展しない。

(2)
```
(IP-MAT (NP-SBJ *pro*)
        (PP (P 如果)
            (IP-ADV (VE 没有)
                    (NP-OB1 (N 基础)
                            (N 研究))))
        (PU ，)
        (ADVP (ADV 就))
        (VE 没有)
        (NP-OB1 (PP (NP (N 科学))
                    (P 的))
                (N 发展))
        (PU 。))
```

3 Penn Historical Corpora 式の解析スキームのメリット

Penn Historical Corpora 式の解析スキームは、機能を表すタグと省略関係を手掛かりとして、述語と述語のとる項との関係が確定できる。これは、Penn Treebank式の解析スキームによっても可能であるが、前者のメリットは主に以下の三点に含まれる。

(a)句の内部構造の一貫性を保つこと。(3)に示すように、一般的には主要部（head）は常に句の要素を投射する(N→NPなど)。句レベルの要素（NPなど）は主要部を直接支配する。それは、X'理論で提唱された中間レベルの欠如ということを意味する。そのため、Penn Historical Corpora 式の解析スキームでは、修飾語（modifiers）も補語（complements）も常に主要部と同じレベルにある。

(3)
```
(XP (Y single-word modifer)
    (YP multi-word modifer)
    (X head)
    (ZP complement))
```

(b) 機能を表すタグが常に節（IP及びCP）にタグ付けされること（第4章を参照）。例えば、IP-MAT (matrix clause)、IP-REL（relative clause）のように、個々のIPはその機能を指すためのタグを持っている。このことが、Penn Historical Corpora 式の解析スキームが高精度の形式的意味表示を自動的に得る研究にもっともふさわしいインプットを提供できると言える最大の理由である。

4 具体例

複文を解析する際に、句と句がどのような関係で組み合わされて文を構成するかということを明らかにする必要がある。句の組み合わせの仕方の一種類として、(4)に示すような連体修飾節が挙げられる。(4)aは英語にも見られるような、通常の関係節による主名詞の修飾の例であり、"机会"（チャンス）は連体修飾節 "之前失去"（この前失った）の目的語として機能しており、2つの句の意味は文の意味表示において、∧で連結される等位構造の関係にある。これに対して(4)bでは、"机会" は "去美国"（アメリカへ行く）の中で文法的役割を持たず、後者は前者の内容を表し、文の意味表示では前者の中に後者が埋め込まれることになる。

(4)a
```
之前    失去 的        机会      又
この前 失う 補文標識 チャンス また
来          了。
やってくる 過去
```
この前失ったチャンスがまたやってきた。

(4)b
```
去    美国      的        机会      又
行く アメリカ 補文標識 チャンス また
来          了。
やってくる 過去
```
アメリカに行くチャンスがまたやってきた。

(4)に関するPenn Treebank式の統語解析木は(5)と(6)である。連体修飾節を伴う名詞句は(5)では二つのCP階層を持っているのに対して、(6)ではCP階層を一つしか持っていない。

(5)
```
(IP (NP-SBJ (CP (WHPP-1 (-NONE- *OP*))
                (CP (IP (NP-SBJ (-NONE- *pro*))
                        (VP (NP-OBJ
                               (-NONE- *T*-1))
                            (ADVP (AD 之前))
                            (VP (VV 失去))))
                    (DEC 的)))
            (NP (NN 机会)))
    (VP (ADVP (AD 又))
        (VP (VV 来)
            (AS 了)))
    (PU 。))
```

(6)
```
(IP (NP-SBJ (CP (IP (NP-SBJ (-NONE- *pro*))
                    (VP (VV 去)
                        (NP-OBJ (NR 美国))))
                (DEC 的))
            (NP (NN 机会)))
    (VP (ADVP (AD 又))
        (VP (VV 来)
            (AS 了)))
    (PU 。))
```

一方、同じく(4)に関するPenn Historical Corpora 式の統語解析木は(7)と(8)である。両方ともCP階層を一つだけ持っているが、CPにその機能を表すタグ（-RELと-THT）が直接与えられている。

(7)
```
(IP-MAT (NP-SBJ (CP-REL (WNP 0)
                        (IP-SUB (NP-OB1 *T*)
                                (NP-SBJ *pro*)
                                (ADVP
                                  (ADV 之前))
                                (VB 失去))
                        (C 的))
                (N 机会))
    (ADVP (ADV 又))
    (VB 来)
    (AXD 了)
    (PU 。))
```

(8)
```
(IP-MAT (NP-SBJ (CP-THT (IP-SUB (NP-SBJ *pro*)
                                (VB 去)
                                (NP-OB1
                                  (NPR 美国)))
                        (C 的))
                (N 机会))
        (ADVP (ADV 又))
        (VB 来)
        (AXD 了)
        (PU 。))
```

即ち、Penn Treebank式の解析スキームでは句の統語木の構造的な違いを作ることによって同じカテゴリーに属する句の機能を区別するのに対して、Penn Historical Corpora 式の解析スキームではより直接的な形式（機能タグの添付）でその違いを示している。そのため、今回句の内部を探索してその構造を解析しなくとも、句の機能タグを見るだけで重要な情報へのアクセスが可能になる。その点で、Penn Historical Corpora 式の解析スキームはPenn Treebank式の解析スキームよりも優れており、述語と述語のとる項との関係の確定だけでなく、句と句の間の関係へのアクセスも容易にしている。

-RELでマークされた(7)の統語構造により形式意味表示として(9)が得られる。“机会” は、補足節の述語 “失去” の取る目的語である（失去机会）とともに主節の述語 “来” の主語にもなっている(机会来)。一方、-THTによって表示された(8)の統語構造は(10)のような形式意味表示を提供する。(9) と同じように “机会” は主節の述語 “来” の主語にはなっているが、(9)と違って補足節の述語 “去” とは特に関係を持っていない。補足節の部分は “机会” の内容としてその中に埋め込まれていることが明らかであろう。

(9) $\exists xye_2e_1(\text{pro:y} = ? \wedge \text{机会}(x) \wedge \text{失去}(e_1, y, x) \wedge \text{之前}(e_1) \wedge past(e_2) \wedge \text{来}(e_2, x) \wedge \text{又}(e_2))$

(10) $\exists xye_2e_1(\text{pro:y} = ? \wedge \text{机会}(x,\text{去}(e_1, y,)) \wedge past(e_2) \wedge \text{来}(e_2, x) \wedge \text{又}(e_2))$

5 統語木の変換

統語コーパスの変換は、tsurgeon tool（Levy and Andrew 2006）を活用したスクリプトの利用および人手による修正の両方によって可能になる。(11)は名詞句の並列関係を表すPenn Treebank式の部分統語木である。これを(13)のようなPenn Historical Corpora 式の木に変換する作業は二つのステップに分けられる。まずはsedスクリプトを使って接続詞を示すスピーチタグCCをPenn Historical Corpora 式のCONJに直す必要がある。次に統語木の構造の変換を行う。(12)に示すtsurgeonスクリプトは、(a)オリジナル木構造の描写及び、(b)実行するアクション、という二つの部分からなっている。以上の二段階を通して、ようやく(13)が得られる。

(11)
```
(NP (NP ...)
    (CC ...)
    (NP ...))
```

(12)
```
(a) NP <1 NP=x < CONJ=y <-1 NP

(b) adjoinF (CNJP @) x
    move y $- x
```

(13)
```
(NP (CONJP (NP ...)
           (CONJ ...))
    (NP ...))
```

以上に示したようにtsurgeonスクリプトは強力で使いやすいが、スクリプトだけではなかなかうまく扱えないような統語木の例もある。(5)にみられた二つのCP階層を持つような関係節構造を変換する場合は、最初のステップとしてまず二番目のCP（CPの下にあるCP）を削除する必要がある。これは、(14)のようなスクリプトを作成し実行すれば簡単に実現できる。

(14)
```
CP < CP=x

excise x x
```

しかし、(14)は処理の過剰一般化の間違いを引き起こす可能性がある。(15)、(16)は、それぞれPenn Treebank式の中国語二重文末助詞文とその統語木をシンプル化したものである(SPは文末助詞を表すスピーチタグのこと)。

(15)
```
現在 差不多   两年   了      嘛。
今   おおよそ 二年間 文末助詞 文末助詞
今までおおよそ二年間になったよね。
```

(16)
```
(CP (CP (IP ...)
        (SP ...))
    (SP ...))
```

(16)は二つのCP階層を持ち明らかに(14)の処理対象になるが、実はその適用は望ましくない処理である。高精度の形式的意味表示を得るためには、二重文末助詞文の場合、二つの文末助詞を同じレベルに置くよりもそれぞれのCP階層に埋め込ませるほうが適切だと思われるからである。(12)のスクリプトの(a)におけるオリジナル木構造の描写の部分をより精密的に指定することによって、ある程度の過剰一般化を避けることが出来る。その前に、まず(16)の中のSPをPにする必要がある。その理由は、(11)、(12)、(13)で示したように、統語木の構造の変換を行う前にsedスクリプトによるスピーチタグの修正（Penn Treebank式からPenn Historical Corpora 式へ）が必要であり、文末助詞の場合は、Penn Treebank式ではSPになるが、Penn Historical Corpora 式では助詞と同じようにPに当たる。次に、スクリプト(14)を(17)に直す。

(17)
```
CP < CP=x !< P

excise x x
```

(17)は「CPの下にCPがあってかつPがないような木構造に対して、CPの下のCPを削除せよ」ということを意味している。これで、(5)と(16)の識別が可能になり、正しい処理が自動的に行えるようになりそうだが、実は(17)もまだ不十分で、すべての状況に対応できたとはいえない。

(18)
```
如果        社会 政策 不    介入 的话,
従属接続詞 社会 政策 否定 介入 文末助詞
```
もし社会政策が介入しないなら、

(19)
```
(CP (ADVP (CS ...))
    (CP (IP ...)
        (XP ...)))
```

(18)は、Penn Treebank式の中国語の条件を表す従属節で、(19)はその統語木をシンプル化したものである(CSは従属接続詞の意味)。更に、(19)のスピーチタグを修正したものは以下の(20)である。

(20)
```
(CP (P ...)
    (CP (IP ...)
        (XP ...)))
```

(20)はCPの下にPがあるため(17)の処理対象には当てはまらないが、実は中国語の条件を表す従属節は(5)の関係節構造と同じように、さらなる木構造変換のため二番目のCP（CPの下にあるCP）をまず削除する必要がある。(20)にも対応できるように、(17)を更に(21)に進化させる必要がある。

(21)
```
CP < CP=x !<-1 P

excise x x
```

(21)は、「CPの下にCPがあって且つ一番右の要素がPではないような木構造に対して、CPの下のCPを削除せよ」ということである。即ちCPの下にあるPに関する条件を(17)よりも厳密的に設定することにより、(5)—関係節構造、(16)—二重文末助詞文、(19)—条件を表す従属節、に全部対応させることが出来た。(21)が有効なのは、中国語の場合、従属接続詞は常に従属節の最初（Pの位置が一番左）に現れるという文法構造上の特徴があるからである。作業を行っていく際に、どのように工夫してもスクリプトだけでコントロールが困難な例も出てくると思われるが、そのようなスクリプトの制御限界を超えることによる間違いの修正は人手に任せる必要がある。

以上ではスクリプトの過剰一般化による問題点とその解決策を論じてきた。統語解析木の形式変換の結果として、コーパスの応用の可能性が広くなり、その汎用性が一段と高まることが期待出来る一方、変換の過程で情報を失う恐れも伴っている。それは、主に統語木の階層を削減する際に起こる問題だと思われ、典型的なものとして複合名詞句の例が考えられる。名詞（N）或いは名詞句（NP）間の関係へのアクセスを保つ手法は、やはりPenn Historical Corpora式の解析システムの要をなす句への機能タグの追加である。第4章にみたように統語構造として含まれた情報を機能タグの中に漏れなく変換する必要があるが、そのためには応用目的に応じて新たな機能タグの作成と追加が必要となる。また、木構造の変換作業は一連のtsurgeonスクリプトの実行であるため、スクリプト同士がお互いに与え合う望ましくない影響を最大限に避けられるように、何百個もあるtsurgeonスクリプトの実行の順番を論理的に設定することも重要だと思われる。以上に挙げたことを考慮しながら作業を行っていくことは大事な課題である。

6 まとめ

この論文は、統語コーパスの重要性およびその作成の困難さから生じる矛盾の克服のために統語コーパス間の転換を提案し、2種類のツリーバンク（Penn TreebankとPenn Historical Corpora）の特徴を紹介した上で、Penn Historical Corpora式の解析スキームが筆者らの応用目的に相応しい理由も述べ、Penn Chinese Treebank (3.0)を対象とする実際の変換例を挙げた。また、統語木の変換に使うtsurgeon toolを活用したスクリプトの紹介を踏まえて、それに伴う問題点を挙げつつ、その解決法を提示した。最後に、これから作業を行っていく際に、予測できる注意点の全体像を描いた。

参考文献

Butler, Alastair and Kei Yoshimoto. 2012. Banking meaning representations from treebanks. *Linguistic Issues in Language Technology - LLT*7(1):1-22.

Levy, Roger and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structure. In *5th International Conference on Language Resources and Evaluation.*

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

Xue, Nianwen and Fei Xia. 2000. The bracketing guidelines for the Penn Chinese Treebank (3.0). Tech. Rep. 00-08, Institute for Research in Cognitive Science, University of Pennsylvania.