山崎直樹(2019).
「古典中国語のテクストをいかに切り分けるか」,
『開篇』37, 111-119.

古典中国語のテクストをいかに切り分けるか

山崎直樹（関西大学）

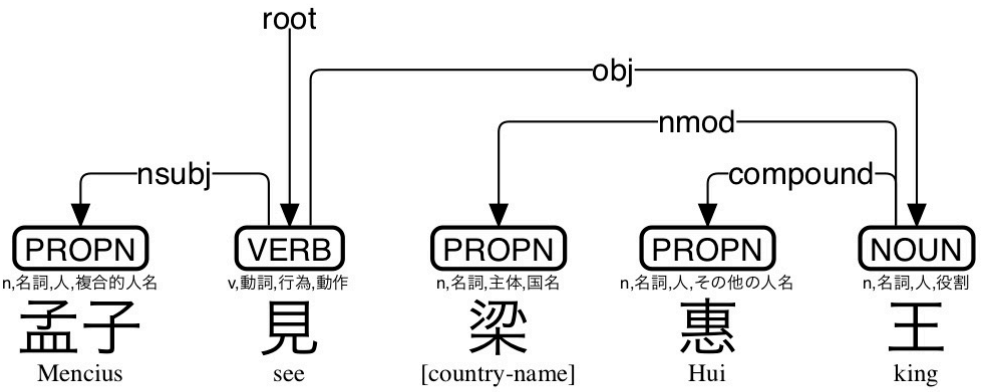
1. この文章の背景と目的

筆者は、現在、「古典中国語の文法解析の自動化」という趣旨のプロジェクトに参加している[1)]。これは簡単にいうと、古典中国語の形態素～文法解析を計算機で行うという試みである。

このプロジェクトは、言語に依存しない形態素解析器である Mecab[2)]を利用して古典中国語を形態素解析することで成果を挙げ（守岡 2008, 2009 を参照）、Mecab で使える形式で階層化した品詞分類を試作し（山崎他 2012 を参照）、これらを利用して、現在は、Universal Dependencies[3)]という枠組（いわゆる「依存文法」を言語普遍的で機械可読な形に形式化したもの）により依存構造解析に取り組んでいる（詳細は、安岡他 2018, 安岡 2018 を参照していただきたい）。

古典中国語解析用に学習をした Mecab と UDPipe[4)]（Universal Dependencies のための解析器）を用いて《孟子》の一節を解析し、安岡孝一（京都大学）の開発による可視化ツール[5)]で表示した例を図 1 に示す。

図 1:解析と可視化の例



[1]この研究は、科研基盤研究(B)「古典漢文形態素コーパスにもとづく動詞の作用域の自動抽出」(17H01835)の援助を受けている。
[2] http://taku910.github.io/mecab/ （2019.2.10 確認）
[3] https://universaldependencies.org/ （2019.2.10 確認）
[4] https://ufal.mff.cuni.cz/udpipe/models （2019.2.10 確認）
[5] http://kanji.zinbun.kyoto-u.ac.jp/~yasuoka/kyodokenkyu/2018-10-26/UDPipeSVG.html （2019.2.10 確認）

この文章では、このプロジェクトの研究に資するために、本来「文」という単位に区切られていない古典中国語のテクストはどのように分割されうるかという可能性を、伝統的な「文」の定義から外れたところで考えてみたい。なぜテクストをより小さな単位に分割する方法を考えないといけないかは、次節で説明する。

2. 問題点

§1 で言及した Universal Dependencies による解析は有望であるが、まだ問題点がある。その 1 つは、文に区切られていない白文のテクストを入力とした場合、それをさらに細かい単位（例えば「文」という単位）に分割する処理に関しては、現時点ではまだ十分ではないことである[6]。

周知のとおり、古典中国語のテクストは、本来、句読点などを付されておらず、文の区切りも示されていない。しかし、依存文法は、「文」という単位の中で、依存関係の最上位にある要素に向かって全てを依存させるという形式化をおこなう（Universal Dependencies では root という節点でこれを示す。図 1 参照）。よって、与えられたテクストを「文」に区切ることが必要なのだが、その処理がまだ十分ではないということである。

3. 「文」の定義

「文」の定義は、意味的な統合性によるものと形式的な特徴によるものがある。現在の自然言語処理では、前者は形式化が困難である。後者は言語によっては比較的簡単である。

例えば、英語の場合、定形節（述語動詞が時制を持つ）は動詞の形態によって明瞭にそれが示され、必ず文法的な主語を伴い、接続詞なしに連続することがない。日本語であれば、文末には述語が現れ、述語を構成する動詞とその接辞はそこで終わるか次に接続するかを派生形で明示する。しかし、古典中国語の場合、これらの特徴を一切持たない。文末にモダリティを表す助詞が現れることもあるが義務的ではないし、それらの助詞には、文中に現れるものもある（主題の提示など）。現に、古典中国語（SVO を基本語順とする）の解析では、動詞の取る目的語がどこで終わるか（≒文がどこで終わるか）を判断することが大きな技術上の難点になっている。

4. 人手による分割

古典の注釈を人手で行なっている場合、どのように「文」に分割しているのか。次の 1a は、『列子集釋』（楊伯峻撰, 中華書局, 1979）「巻第五湯問篇」の有名な「愚行移山」の部分である。1b は、『列子』（小林信明, 明治書院, 1967）の相当する箇所である。

---

6 いっぽうで、「文」に区切った形で入力を与えられれば解析はかなりの精度を示す（安岡 2019a, 2019b）。

1a. 北山愚公者，年且九十，面山而居。懲山北之塞，出入之迂也，聚室而謀，曰：「吾與汝畢力平險，指通豫南，達于漢陰，可乎？」

1b. 北山愚公者、年且九十。面山而居、懲山北之塞、出入之迂也。聚室而謀曰、吾與汝畢力平險、指通豫南、達于漢陰。可乎。

句レベルの区切りはほぼ一致しているが、どこに「。」を打つか（≒どこまでを一文とするか）は恣意的であるように見える。ほんとうは恣意的なのではなく、形式化しにくいノウハウ、経験的な感覚などによった判断に基づいているのであろうが、この部分の判断を計算機におこなわせるのは難しかろう。

前節で述べたとおり、古典中国語のテクストを近代言語学の定義による「文」に分割するのは、現時点では計算機にとってはなかなか難しい。計算機が判定可能にするためには、これまでの「文」の定義とは構造的な性質が（多少）異なっても、もっと形式的に明確な指標を用いたわかりやすいテクストの分割の方法を考えるのも 1 つの方向であろうと思う。

5. 主題連鎖による分割

（以下に述べることは、「古典中国語の文法解析の自動化」プロジェクトの総意ではなく、自然言語処理の技術的な側面には疎い筆者個人の素人考えであることをお断りしておく）

私見によれば、「形式的に明確な指標を用いた、わかりやすい」方法は、主題が作る連鎖 (Topic Chain) による分割である。

古典中国語は、ある節の主語が後続する節と同一指示である場合、後続する節の主語は省略されることがふつうである。この「主語が省略された節」は先行する節の主語に連鎖上にぶらさがっていると考え、これが終わるところ（次の主題が出現する直前）までを、1 つの単位とみなす方法である。

次は、前節で見た「愚公移山」を主題連鎖で分割した例である[7]。一見すると分かるが、現代の代表的な標点本（中華書局本など）の分割のしかたと大差はない。この切りかたは人の感覚とも合致するのである。太字で示した部分が主題と目される項目で、「。」があるところが連鎖の終わりである。「φ」は主語があるのであればそこに現れるであろうという位置である。「，」は読みやすさを考慮して筆者が挿入した。

2. **北山**愚公者年且九十，φ 面山，而 φ 居，φ 懲山北之塞，φ 出入之迂也，φ 聚室而 φ 謀，φ 曰。

7 以下で使用する古典中国語のテクストは、『列子集釋』（中華書局, 1979）、『史記』（中華書局, 2013）、『十八史略』（新釈漢文大系, 明治書院, 1967）所載のものから、標点などを取り去って使用した。

吾與汝畢力平險，φ 指通豫南，φ 達于漢陰，φ 可乎。

以下、いくつか主題連鎖で切った例を見てみる。

3. **孔子**生魯昌平乡陬邑。
   先宋人也，φ 曰孔防叔。
   防叔生伯夏。
   伯夏生叔梁紇。
   紇与顔氏女野合，而 φ 生孔子，φ 祷于尼丘 φ 得孔子。（史記・孔子世家）
4. **秦始皇帝者**秦莊襄王子也。
   莊襄王為秦質子於趙，φ 見呂不韋姬，φ 悦而 φ 取之，φ 生始皇，φ 以秦昭王四十八年正月生於邯鄲。（史記・始皇本紀）
5. **蘇秦者**東周雒陽人也，φ 東事師於齊，而 φ 習之於鬼谷先生，φ 出遊數歲，φ 大困而歸。
   兄弟嫂妹妻妾竊皆笑之，φ 曰
   周人之俗，φ 治產業，φ 力工商，φ 逐什二以為務。
   今**子**釋本，而 φ 事口舌，φ 困，φ 不亦宜乎。
   蘇秦聞之，而 φ 慚，φ 自傷，φ 乃閉室，φ 不出，φ 出其書，φ 偏觀之，φ 曰。
   夫**士**業已屈首，φ 受書，而 φ 不能以取尊榮，φ 雖多，亦 φ 奚以為。於是 φ 得周書陰符，φ 伏，而 φ 讀之，期年 φ 以出揣摩，φ 曰。
   此可以說當世之君矣，φ 求說周顯王。
   顯王左右素習知蘇秦，φ 皆少之，φ 弗信。（史記・蘇秦列伝）
6. **孝文王**立，三日而 φ 薨。
   楚立。
   是爲莊襄王，φ 四年薨。
   政生十三歲矣，φ 遂立，φ 爲王。
   母爲太后。
   不韋在莊襄王時，φ 已爲秦相國，至是 φ 封文信侯。
   太后復與不韋通。（十八史略・秦）
7. **王**既長。
   不韋事覺自殺。
   太后廢，φ 處別宮。
   茅焦諫。
   母子乃復如初。（十八史略・秦）

上の例は、筆者が、主題連鎖と見なしてよいと思われる構造に、人手で切り分けたもので

ある。ただし、この切り分けはあくまで形式から判断したものであって、上の例の中では、（ここが重要なのであるが）「文頭」にある主題と φ を仮定した項が必ずしも同一指示でない連鎖も存在する。よって、この主題連鎖による切り分けを機械的に行えるようになった場合でも、それをそのままテクストの意味解釈に用いることはできない。しかし、形式的な特徴によりテクストをより細かい単位に分割するということと主題連鎖を意味解釈に使うことを別の過程だと考えれば問題はない（その場合、個々の連鎖の内部での意味解釈は別の処理でおこなう必要がある）。

6. 主題の見つけかた

主題連鎖による分割が有効であるとして、機械的にこれをおこなうためには、「主題」を形式的に見つける処理過程が必要になる。

人の事績を記したようなテクストでは、最も主題になりやすいのは「人」であろうから、人を指す名詞（人名を含む）を見つけられるかが 1 つのポイントである。上述の図 1 で見たように、我々の解析システムでは、人を表すような形態素を含んでいれば、その名詞句が人を示すものであることを、あるていど判断できる。「人を指す名詞（句）」が動詞性を持つ形態素に支配されていなければ（＝他動詞や前置詞の目的語になっていなければ）主題と見なす」という処理により、主題は多くの場合、発見可能であると思われる。

また、安岡他(2014)で示したように、Mecab を使った形態素解析は、辞書の充実により、地名に関してはあるていど正確に切り出せるようになっている。問題は固有名詞のみからなっている（地名、役職名、尊称、親族呼称などを含まない）人名だが、これは辞書の拡充を待つしかない。

もちろん、このような簡単な処理で全ての課題が解決するわけではない。以下では、問題となりそうな点をいくつかピックアップする。

7. 主題と見紛う項目

人の事績を記したテクストでは「人」が主題になりやすいが、「人」を示す語を正確に切り出すのは、現時点の我々のシステムでは難しい。そうなると、動詞に支配されていない位置（つまり主題が現れうる位置）に「人でない名詞」が現れたときが問題となる。それが主題と見なしてよい名詞句であればかまわないのだが、必ずしもそうとばかりは限らないであろう。また、さらに、人を表す主題よりも前に（＝さらに卓立性の高い位置に）人を指すのではない名詞句が現れたときも問題になる。一般的にいって、主題が複数現れると、それら相互の関係や、後続する述語との関係の判断が難しいからである。

8. **今，子**釋本，而 φ 事口舌，φ 困，φ 不亦宜乎。（史記・蘇秦列伝）
9. **昔，穆公**取由餘於戎，φ 得百里奚於宛，φ 迎蹇叔於宋，φ 求丕豹公孫枝於晉，φ 并

國二十，φ 遂覇西戎（十八史略・秦）

10. **是歳，季武子**卒。（史記・孔子世家）

11. **昭襄王時，孝文王柱**爲太子，φ 有庶子楚，φ 爲質于趙。（十八史略・秦）

8 の例では、動詞に束縛されていない主題位置の要素は“今”と“子”である。意味的に考えれば、人を表す“子”が連鎖を作っていることは明らかなのであるが、より卓立性の高い位置にあるのは“今”である。9 の“昔”と“穆公”、10 の“是歳”と“季武子”、11 の“昭襄王時”と“孝文王柱”も同じである。

ただ幸いなことに、これらはみな「時」と関係ある表現であり、なおかつそれを示す形態素を内に含む。機械的な判定は難しくない。

12. **由是，孔子**疑其父墓處。（史記・孔子世家）

13. **然後**，φ 往合葬於防焉。（史記・孔子世家）

14. **於是**，φ 大索逐客。（十八史略・秦）

12 の例は、人を表す主題よりも卓立性の高い位置に他の語句がある例であり、他の例は、人を表す主題がなくて、別の語句が主題の位置にある例である。一瞥してわかるとおり、これらはいわゆる「接続詞」的なフレーズであり、このように用いられる表現は限定されている。機械的な判定は難しくない。

8. 典型的な連鎖を作らない文型

主題連鎖を作りやすいのは、ある主題に関してその属性の描写を連続させるような構造か、ある動作者を主題にして、「××が…を〜して、…を〜して」のように、典型的な他動詞句が続くような構造である。

形式的な特徴によりテクストをより細かい単位に分割するということと、主題連鎖を意味解釈に使うことを別のものだと考えるのであれば、他動詞文の連鎖を主題連鎖で処理することに問題はないが、ここでは、参考のため、後続する動詞句の主語位置の名詞句と主題とを同一指示と解釈することが誤った解釈を引き起こす例を見てみたい。

15. **鄰人京城氏之孀妻**有遺男，φ 始齕跳往助之。（列子・湯問）

16. **不韋**因納邯鄲美姫，φ_1 有娠而獻于楚，φ_2 生政，φ_3 實呂氏。（十八史略・秦）

17. **莊襄王**為秦質子於趙，φ_1 見呂不韋姫，φ_2 悦，而 φ_3 取之，φ_4 生始皇，φ_5 以秦昭王四十八年正月，φ_6 生於邯鄲。（史記・始皇本紀）

15 の例は、後続する φ を支配するのは主題位置にある“鄰人京城氏之孀妻”ではなく、動詞“有”の目的語の“遺男”である。16 の例で、ϕ_1 と ϕ_2 を支配するのは“邯鄲美姫”

で、最後の ϕ_3 は“政”が支配する。つまり、どれも主題位置にある「不韋」の支配ではない。17 の例では、φ_1 から φ_3 までが主題位置の“莊襄王”の支配で、φ_4 は“呂不韋姬”の支配、φ_5 と φ_6 は“始皇”の支配であるが、“始皇”はこの一連の連鎖の中では主題位置に現れていない。

実は、このように、談話の中に新しく導入された項目が、主題位置ではなく目的語位置にあるにも関わらず焦点を担い、その後の空所をコントロールするという現象は、存在や出現を表す文のような非典型的な他動詞文の場合（上の例はみな存在や出現と関連する表現である）、珍しくはない。現代中国語でも頻繁に観察される（山崎 1995 を参照）。

9. 主題の「入れ子」

主題の位置に現れる動詞に支配されていない項目が、複数、現れうることは上述した。ここでは、主題が作る連鎖が「入れ子」になっている現象を見てみたい。

表 1: 主題連鎖が入れ子になっている例（史記・孔子世家）

	接続詞的要素	主題 1	主題 2	述部
01		孔子		貧
02		Φ		且賤
03		Φ		及長
04		Φ		嘗為季氏史
05			料量	平
06		Φ		嘗為司職吏
07	而		畜	蕃息
08	由是	Φ		為司空
09		Φ		已而去魯
10		Φ		斥乎齊
11		Φ		逐乎宋衛
12		Φ		困於陳蔡之間
13	於是	Φ		反魯
14		孔子		長九尺有六寸
15		Φ	人	皆謂之長人而異之
16		Φ	魯	復善待
17	由是	Φ		反魯

表 1 に示したテクストでは、05 の“料量”、07 の“畜”、15 の“人”、16 の“魯”などの

主題（動詞性の語に支配されていない項目）を跳び越え、“孔子”が連鎖を成していることが見て取れる。

このテクストの内容を熟知する「人」であれば、意味的に見て、“孔子”の卓立性が他の主題より高いことを違和感なく理解できるであろうが、これを機械処理で判定するためには、まだまだ解決しなければならない課題が多そうである。

10. まとめ

主題連鎖によるテクストの分割も実現までにはまだ課題が多いと思われるが、実現すれば、解析の精度も上がることが期待される。次は、『十八史略』「秦」の任意の一部である。18a は区切りの無い白文、18b はそれを入力として、我々のシステムで解析した場合、どこで文を区切り（「 / 」で示した）、どの項目を root と判断したか（太字で示した）である。18c は手作業で主題連鎖を切り出し、その連鎖ごとに入力を分けて解析した場合である。

18a. 孝文王立三日而薨楚立是爲莊襄王四年薨政生十三歳矣遂立爲王母爲太后不韋在莊襄王時已爲秦相國至是封文信侯太后復與不韋通王既長不韋事覺自殺太后廢處別宮茅焦諫母子乃復如初

18b. 孝文王立三日而薨楚立是爲**莊**襄王 / 四年**薨**政 / 生十三**歳**矣 / 遂**立**爲 / 王母**爲**太后不韋在莊襄王 / 時已**爲** / 秦相國**至** / 是**封**文 / 信侯太后**復**與不韋通 / 王既**長**不韋事覺自殺 / 太后**廢**處別宮 / 茅焦**諫**母子 / 乃復**如**初

18c. 01: 孝文王**立**三日而薨
02: 楚**立**
03: 是**爲**莊襄王 / 四年**薨**
04: 政**生十**三歳矣 / 遂立爲**王**
05: 母爲**太后**
06: 不韋**在**莊襄王 / 時已**爲** / 秦相國**至** / 是**封**文信侯
07: 太后**復**與不韋通
08: 王既**長**
09: 不韋**事**覺自殺
10: 太后**廢**處別宮
11: 茅焦**諫**
12: 母子乃**復**如初

主題連鎖で切り分けてからのほうが、より細かい単位への分割において、root の決定においても精度が高くなっている印象を受ける。

なお、04 の後半と 05 で動詞の後の名詞が root になっているのは、「コピュラ（繋辞）を使った文では root はコピュラの後の名詞とする」という Universal Dependencies の規則

に拠っている。03 で“爲”を root とするのがむしろ規則に沿っていない。このあたりはまだ安定していないところである。

参考文献

守岡知彦(2008).「MeCab を用いた古典中国語の形態素解析の試み」『情報処理学会研究報告』Vol.2008-CH-79, pp.17–22.

守岡知彦(2009).「MeCab を用いた古典中国語形態素解析器の改良」『情報処理学会研究報告』Vol.2009-CH-84, No.3, pp.1–5.

安岡孝一(2018).「古典中国語（漢文）の依存文法解析と直接構成素解析」『漢字文献情報処理研究』Vol.18, pp.55-61.

安岡孝一(2019a).『古典中国語 Universal Dependencies で読む『孟子』』（センター研究年報 2018 別冊）, 京都大学人文科学研究所漢字情報研究センター.

安岡孝一(2019b).「四書を学んだ MeCab + UDPipe はセンター試験の漢文を読めるのか」『東洋学へのコンピュータ利用』第 30 回研究セミナー, 京都大学, 2019/3/8.

安岡孝一, 守岡知彦, ウィッテルン・クリスティアン, 山崎直樹, 二階堂善弘, 鈴木慎吾(2014).「古典中国語形態素解析による地名の自動抽出」『人文科学とコンピュータシンポジウム（じんもんこん）2014 論文集』pp.63-68.

安岡孝一, ウィッテルン・クリスティアン, 守岡知彦, 池田巧, 山崎直樹, 二階堂善弘, 鈴木慎吾, 師茂樹(2018).「古典中国語（漢文）の形態素解析とその応用」『情報処理学会論文誌』Vol.59, No.2, pp.323-331.

山崎直樹(1995).「物語における新規項目の導入と文型の選択」『中国語学』242, pp.115-122.

山崎直樹, 守岡知彦, 安岡孝一(2012).「古典中国語形態素解析ための品詞体系再構築」『人文科学とコンピュータシンポジウム（じんもんこん）2012 論文集』pp.39-46.