四書を学んだコンピュータはセンター試験の漢文を読めるのか

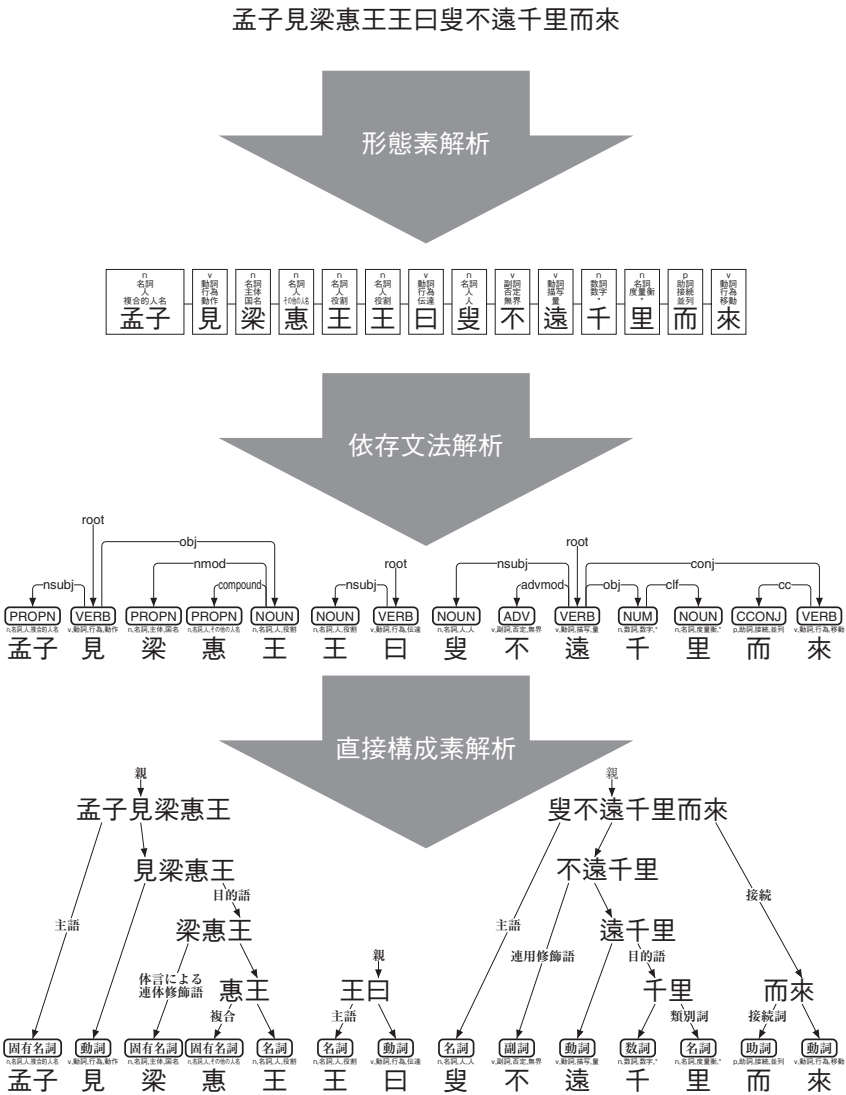
安岡孝一

コンピュータによる言語処理という観点から見ると、古典中国語（漢文）の白文というのは、かなり厄介なシロモノである。単語と単語の間に区切りがない。文と文の間にも区切りがない。漢字がズラズラと切れ目なく並ぶだけ。こんなもの、どうやって読めばいいのか。

共同研究班「東アジア古典文献コーパスの実証研究」では、古典中国語における文法解析の自動化に全力で取り組んでいる。漢文の白文に対し、形態素解析・依存文法解析・直接構成素解析を順におこなうことで、白文の統語構造が解析可能となる（次ページ図）というのが、われわれの見通しである。形態素解析によって、単語切りをおこなうと同時に、各単語の品詞を得る。依存文法解析によって、単語と単語の間の係り受け関係を解析すると同時に、文の切れ目を得る。直接構成素解析によって、各文の統語構造を解析木の形で得る。われわれは、これまでに古典中国語の形態素解析をほぼ完成し、現在は依存文法解析の手法開発に精力を集中している。その一方、直接構成素解析については、残念ながら、本稿執筆時点（二〇一九年二月）ではアイデアどまりである。

これまでわれわれは、古典中国語の形態素解析に、MeCab（以下M）という形態素解析エンジンを使ってきた。Mは、もともとは日本語向けの形態素解析エンジンだったが、言語、辞書、コーパスに依存しない汎用的な設計がなされており、古典中国語への適用も可能である。ただしMは、あくまで形態素解析エンジンであり、依存文法解析はおこなえない。これに対し、UDPipe（以下U）という形態素・依存文法解析エンジンと、今年になってStanfordNLP（以下S）という形態素・依存文法解析エンジンが現れたことから、これらの性能比較をおこなうことにした。Uは、チェコ語の係り受け解析エンジンが土台となっているが、多言語化をおこなった際に単語切り・文切り・品詞付与の機能が追加され、形態素解析と依存文法解析の両方を兼ね備えたものとなっている。Sは、英語の品詞付与・係り受け解析エンジンが土台となっており、多言語化を行った際に文切り・単語切りの機能が追加されたものである。

われわれの手元には、四書（孟子・論語・大学・中庸）の形態素コーパスと依存文法コーパスがあり、こ

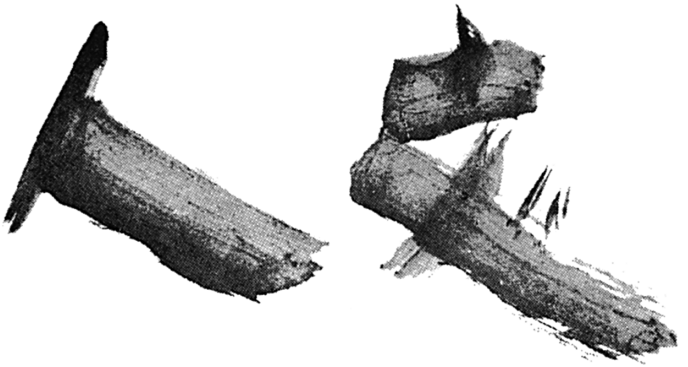

孟子見梁惠王王曰叟不遠千里而來
形態素解析
n 名詞 人 複合的人名 孟子
v 動詞 行為 動作 見
n 名詞 主体 固名 梁
n 名詞 人 その他人名 惠
n 名詞 人 役割 王
n 名詞 人 役割 王
v 動詞 行為 伝達 曰
n 名詞 人 人 叟
v 副詞 否定 無界 不
v 動詞 描写 量 遠
n 数詞 数字 千
n 名詞 度量衡 里
p 助詞 接続 並列 而
v 動詞 行為 移動 來
依存文法解析
root
obj
nmod
nsubj
compound
PROPN n.名詞.人.複合的人名 孟子
VERB v.動詞.行為.動作 見
PROPN n.名詞.主体.固名 梁
PROPN n.名詞.人.その他の人名 惠
NOUN n.名詞.人.役割 王
root
nsubj
NOUN n.名詞.人.役割 王
VERB v.動詞.行為.伝達 曰
root
nsubj
NOUN n.名詞.人.人 叟
advmod
ADV v.副詞.否定.無界 不
VERB v.動詞.描写.量 遠
obj
NUM n.数詞.数字.* 千
clf
NOUN n.名詞.度量衡.* 里
conj
CCONJ p.助詞.接続.並列 而
cc
VERB v.動詞.行為.移動 來
直接構成素解析
親
孟子見梁惠王
主語
見梁惠王
目的語
梁惠王
体言による連体修飾語
惠王
複合
固有名詞 n.名詞.人.複合的人名 孟子
動詞 v.動詞.行為.動作 見
固有名詞 n.名詞.主体.固名 梁
固有名詞 n.名詞.人.その他の人名 惠
名詞 n.名詞.人.役割 王
親
王曰
主語
名詞 n.名詞.人.役割 王
動詞 v.動詞.行為.伝達 曰
親
叟不遠千里而來
主語
不遠千里
連用修飾語
遠千里
目的語
千里
類別詞
接続
而來
接続詞
名詞 n.名詞.人.人 叟
副詞 v.副詞.否定.無界 不
動詞 v.動詞.描写.量 遠
数詞 n.数詞.数字.* 千
名詞 n.名詞.度量衡.* 里
助詞 p.助詞.接続.並列 而
動詞 v.動詞.行為.移動 來

れらを学習用コーパスとして利用できる。そこでⓂⓊⓈに、それぞれ四書を学習させ、古典中国語の形態素解析・依存文法解析モデルを構成した。このモデルに、四書以外の白文を解析させることで、性能比較をおこなえるわけだ。ただし、Ⓜの形態素辞書は、四書に出てくる単語（形態素）に加え、大量の固有名詞も収録していることから、それと同じ単語をⓊの外部辞書にも追加した。一方、Ⓢには外部辞書を接続する方法が（公開されてい）なかったため、Ⓢに関しては実験条件を揃えきれなかった。

実験対象は、大学入試センター試験『国語』の本試験から、第4問の漢文本文を準備した。評価指標は、MLAS（Morphology-aware Labeled Attachment Score）と呼ばれる百点満点の評価値を用いて、品詞付与と係り受けの正確さを両方同時に評価することとした。ただし、センター漢文は常用漢字で書かれており、われわれの四書コーパスは（いわゆる）康熙字典体で記述していることから、Ⓜの形態素辞書とⓊの外部辞書に、常用漢字の単語も収録した。これらの準備の後、二〇一九年のセンター漢文を白文化（句読点や返り点などを除去）し、Ⓤに読ませてみたところ、解析結果の評価値は三九点という、かなり低い結果となった。同じセンター漢文をⓈに読ませてみたところ、二八点という、やはり低い結果となった。読めていない。Ⓜ＋Ⓤ（形態素解析をⓂで、依存文法解析をⓊでおこなう）では三九点、Ⓜ＋Ⓢでは三六点だった。どうしてこんなに評価値が低いのか。

各モデルの解析結果を手作業で検討してみたところ、単語切り・品詞付与・係り受け解析・文切りの四つの機能のうち、文切りが極めてダメダメであることが判明した。ⓊやⓈの文切りは、機能としては先進的なものの、まだまだ古典中国語には追随しきれていないようなのだ。この仮説を検証するため、白文化の際に句読点を除去せず、文切りを事前に与えるという手法で、単語切り・品詞付与・係り受け解析の評価をおこなってみることにした。そうしたところ、二〇一九年のセンター漢文に対し、Ⓤは六七点、Ⓢは三八点、Ⓜ＋Ⓤは六三点、Ⓜ＋Ⓢは五五点となった。文切りを事前に与えることで、いずれも改善が見られ、特にⓊは、まあまあ及第点となっている。

二〇一八年以前のセンター漢文に対しても、凸凹はあるものの、ほぼ同様の傾向が見られた。これを裏返して考えると、古典中国語の依存文法解析のうち、文切りの自動化に関しては、まだまだ研究の余地があるということだ。今後のわれわれの研究の進展に期待されたい。



第六六号

2019

京都大学人文科学研究所

ISSN 0389－147X