4T-06

人間の漸進的言語処理能力の分析

後藤　亮†　　大野　誠寛†　　松原　茂樹‡

東京電機大学未来科学部†　　名古屋大学情報連携統括本部‡

1　はじめに

入力と同時的に処理を進める言語処理システムに対して構文情報を提供するために，解析処理を漸進的に進めていく係り受け解析技術が開発されてきた[1, 2].

一方，人間の言語理解過程には漸進性があることが知られており（例えば，[3]），その言語処理過程を定量的に分析することができれば，漸進的係り受け解析器の性能を向上させるための知見が得られる可能性がある.

そこで本稿では，複数人による言語処理過程を表出したデータを定量的に分析することにより，人間の漸進的係り受け解析能力や入力予測能力の一端を明らかにする.

2　分析に使用したデータ

本研究では，文献[4]で構築されたデータを更に増築したものを使用して分析する．本節ではまず，文献[4]で構築されたデータのアノテーション情報について解説し，次に，その増築について述べる.

2.1 人間の言語処理過程のアノテーション

人間の言語処理過程を表出する方法は様々考えられるが，文献[4]では，漸進的係り受け解析器の出力を模する形を採用している．以下では，文献[4]が参照する解析器[1]と[2]の出力構造，及び，文献[4]でのアノテーション情報を解説する.

文献[1]の解析器では，文の入力途中において，係り先が既入力の文節については係り先文節との依存関係を，また，係り先が未入力の文節については既入力のどの文節とも依存関係にないことを明示した係り受け構造を出力する．その係り受け構造の例を図 1 に示す．この例は，文「私は米国から日本に帰ってきてサッカーのワールドカップを見ました」のうち，文節「サッカーの」まで入力された段階で出力する係り受け構造を示している.

文献[2]の解析器では，文献[1]の係り受け構造を拡張し，係り先が未入力の文節が複数ある場合は，それらの係り先が同一か否かを明示した係り受け構造を出力する．図2に，図1と同じ入力において出力される文献[2]の係り受け構造を示す.

文献[4]では，上記の解析器の出力構造に基づいて，文節が入力されるごとに，図2の形式の係り受け構造の解析と，その構造の中の未入力係り先文節（例えば図2のaとb）の文字列の予測を人間に行わせ，その情報をアノテーションしている.

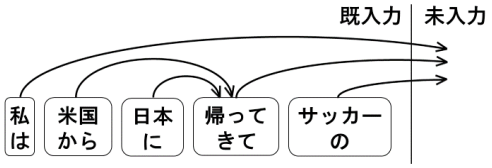

図1 文節間の依存・非依存を明示した係り受け構造[1]

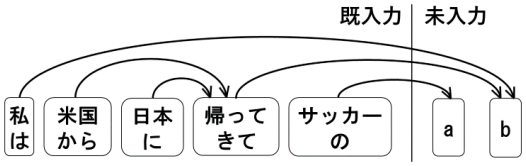

図 2 未入力文節との依存関係を明示した係り受け構造[2]

2.2 データの増築

著者らは文献[4]において，京大テキストコーパスVersion4.0に含まれる毎日新聞記事の2,502文に対して作業者1名によるアノテーションを実施していた．その後，作業者を1名増やし，3,639文に対するアノテーションを作業者2名にそれぞれ実施させ，データの増築を行っており，本研究では，この増築したデータを使用する.

3　人間の漸進的言語処理能力の分析

人間の漸進的言語処理に関する能力の一端を明らかにするため，本節では，係り受け構造の解析と文字列の入力予測の二つに分けて分析する.

3.1　係り受けに関する分析

作業者の係り受け解析能力に関して以下の 3 つの観点で評価する.

- **文単位解析**：文末文節が提示された後に 1 文全体に対して作業者が付与した係り受け構造のみを対象として，正解（京大テキストコーパス上）の係り受け構造とどの程度一致しているかを評価する．これは，文全体の文節列が一度に入力され，その係り受け構造を出力するという一般的な係り受け解析を人間が施した結果の精度に相当する.
- **文節単位解析 A**：作業者が付与した係り受け構造を図 1 に示す形式の係り受け構造とみなして評価する．すなわち，係り先が既入力である文節については係り先文節との依存関係を，係り先が未入力である文節については係り先が未入力であるか否かが正解と一致しているか否かを評価する.
- **文節単位解析 B**：作業者が付与した係り受け構造を図 2 に示す形式の係り受け構造とみなして評価する．なお，作業者が付与した係り受け構造は，係り先が未入力である文節について，その係り先文節を具体的に決めるわけではないため，正解と一致するかを単純には判定できない．そのため，アノテーション結果と正解を比較し，一致する係り受け関係の数が最多となるように，正解とアノテーション結果の係り先文節を動的計画法により対応付け，一致した係り受け関係の数を評価した.

Analysis of Human Capabilities on Incremental Language Processing

Akira Goto[†, a)], Tomohiro Ohno[†, b)], Shigeki Matsubara[‡]

† School of Science and Technology for Future Life, Tokyo Denki University.

‡ Information and Communications, Nagoya University.

a) 16fi048@ms.dandai.ac.jp

b) ohno@mail.dendai.ac.jp

表 1 作業者 X，Y の係り受け解析の正解率

	作業者 X		作業者 Y	
解析名	係り受け正解率	文正解率	係り受け正解率	文正解率
文単位	94.7% (34,875/36,824)	63.3% (2,305/3,639)	95.0% (34,999/36,824)	65.4% (2,380/3,639)
文節単位 A	94.5% (247,957/262,426)	54.6% (1,986/3,639)	94.2% (247,145/262,426)	48.9% (1,779/3,639)
文節単位 B	91.8% (240,893/262,426)	31.5% (1,145/3,639)	89.6% (235,166/262,426)	18.6% (678/3,639)

表 2 係り先が既入力である場合と未入力である場合に分けて評価した結果

		作業者 X			作業者 Y		
		再現率	適合率	F 値	再現率	適合率	F 値
文節単位 A	既入力	95.8% (174,424/182,143)	94.0% (174,424/185,608)	94.9	95.3% (173,556/182,143)	94.3% (173,556/184,015)	94.8
	未入力	95,9% (73,533/76,644)	95.7% (73,533/76,818)	95.8	96.0% (73,589/76,644)	93.9% (73,589/78,411)	94.9
文節単位 B	未入力	86.7% (66,469/76,644)	86.5% (66,469/76,818)	86.6	80.4% (61,610/76,644)	78.6% (61,610/78,411)	79.5

表 3 作業者 X，Y の文字列の入力予測精度

	作業者 X		作業者 Y	
	完全一致	部分一致	完全一致	部分一致
再現率	5.7% (2,917/50,847)	12.5% (6,372/50,847)	3.6% (1,838/50,847)	12.6% (6,412/50,847)
適合率	12.8% (2,917/22,723)	28.0% (6,372/22,723)	6.7% (1,838/27,512)	23.3% (6,412/27,512)

作業者 X，Y それぞれの評価結果を表 1 に示す．なお文正解率は，各観点において 1 文に対するすべてのアノテーション結果が正解と完全に一致している文の割合である．

また，文節単位解析A，Bに関して，係り受け解析精度を，係り先が既入力である場合と未入力である場合に分けて評価した．その結果を表2に示す．なお，文節単位解析Bの「既入力」の値は文節単位解析Aの「既入力」の値と同じになる．

評価結果から，文節単位解析 B において，係り先が未入力の場合の解析は最も難しいタスクであることがわかる．また，各観点での評価結果の差分に着目すると，文単位解析と文節単位解析 A の差より，文節単位解析 A と文節単位解析 B の差が大きいことがわかる．これらは，人間にとって，係り先が未入力であることを示すことはそれほど難しくないが，係り元と未入力文節との間の関係を示すことは難しいタスクであることを示唆している．

次に，作業者XとYのアノテーションの一致率を調査した．全体に対する，両者が共に正解した係り受け関係の割合を算出すると，文単位解析で92.7% (34,123/36,824)，文節単位解析Aで91.7% (240,726/262,426)，文節単位解析Bで85.9% (225,519/262,426)となった．文単位解析，文節単位解析A，文節単位解析Bの順に，その割合は低くなっており，特に，文節単位解析Bの割合が急激に低下している．このことは，係り元と未入力文節との関係を示すことは人間の間で揺れが大きく難しいことを示唆している．

3.2 文字列予測に関する分析

作業者が未入力文節の文字列をどの程度正しく予測できているのかを評価した．表 3 に作業者 X，Y それぞれの予測した文字列の再現率，適合率を示す．再現率は，正解の係り受け構造において，未入力の係り先となる文節 50,847 個のうち，作業者が文字列を正しく予測した場合を意味する．適合率は，それぞれの作業者が何らかの文字列を予測した未入力係り先文節のうち，正しく文字列を予測した場合を意味する．なお，作業者が予測した文字列が正解の文字列と完全に一致している場合のみ正解とした場合と，部分的に一致している場合でも正解とした場合に分けて評価している．この結果から，未入力の係り先文節の文字列を予測することは，人間にとって難しいタスクであるが，全く予測が出来ないわけではないことがわかる．

4 おわりに

本稿では，複数の人間の言語処理過程を表出した大規模データに対して定量的な分析を実施し，人間の漸進的係り受け解析能力や入力予測能力の一端を明らかにした．今後は，より詳細かつ多様な分析を実施し，そこで得られた知見に基づき漸進的係り受け解析器の改善を進める予定である．

謝辞 本研究は，一部，科学研究費補助金基盤研究（B）No. 26280082 及び（C）No. 16K00300 により実施した．

参考文献

[1] 大野，松原，“文節間の依存・非依存を同定する漸進的係り受け解析，” 信学論， J98-D(4), pp. 709-718, 2015.

[2] 相津ら，“漸進的係り受け解析における未入力文節との構文的関係の同定，” 情報処理学会第82回全国大会講演論文集，2020. （発表予定）

[3] G. T.M. Altmann and M. J. Steedman, “Interaction with context during human sentence processing,” Cognition, 30, pp. 191–238, 1988.

[4] 大野，松原，“漸進的係り受け解析の出力構造 –人間の文解析過程のアノテーション–，” 言語処理学会第 22 回年次大会発表論文集，pp. 457-460, 2016.