Bulletin of Aichi Univ. of Education, 63（Humanities and Social Sciences）, pp. 119 − 124, March, 2014

係り受け解析を用いた難読文の抽出とその対策についての一考察

伊藤 俊一

情報教育講座

Detection of Inappropriate Sentence by Using Japanese Dependency Structure Analysis

Toshikazu ITO

Department of Information Sciences, Aichi University of Education, Kariya 448-8542, Japan

1. はじめに

言語は，人間が情報を伝えるための主要な道具である。言語によって情報を正しく伝え，それを相手に正しく理解させるためには，言語の発信者と受信者の双方に高度な情報処理能力が備わっている必要がある。しかし，言語による伝達においては，しばしば情報が正しく伝わらない事態が生じることがある。

日本語文において文意が正しく伝わらない事態を生じさせる状況の1つとしては，正しい係り受け構造の解析が困難な状況が考えられる。本研究では，特に係り受け構造の解析が難しい文に焦点を当てて，そのような文が生じる原因とその対策について考察することにする。

同じ文に対して複数の読み手の解釈が異なる場合，その文は曖昧で読みにくい文であると言える。本研究では，複数の読み手が解析した係り受け構造に違いが生じる文を大量に収集することにより，違いが生じる文に共通する特徴を見つけることを目的とする。そして，係り受け構造の解析が困難なために文が読みにくくなるような状況を回避するための対策について考察していく。

複数の読み手として，本研究では日本語係り受け解析器1種類と人間1名を用いることにする。人間が係り受け構造の解析を行うときには，書き手の意図や文脈を読み取ることによって，本来は係り受け構造の解析が困難であるような文に対してもある程度正しく解析できてしまう場合がある。一方で，現状では係り受け解析器は人間と比べると意図や文脈を読み取る能力が低いため，係り受け構造の解析が難しい文であればあるほど，単純にそれに比例して不適切な解析してしまう可能性も高まると考えられる。そのことを利用して，係り受け解析器と人間がそれぞれに解析した係り受け構造に違いが生じる文を特定することによって，読みにくい状況が生じている文を見つけ出すという手法を本研究では用いることにする。

したがって，本研究では，係り受け解析器と人間の解析結果が異なる場合には，人間の解析結果を正しいもの，係り受け解析器の解析結果を誤ったものとみなして分析を進めることにする。

2. 方法

材料：

大学一年生20名が授業「レポートライティング」の課題としてそれぞれ400字詰め原稿用紙3枚に手書きで作成したレポートの初稿を対象とする。

実験参加者：

授業「レポートライティング」を担当する大学教員1名，および，日本語係り受け解析器Cabocha（工藤・松本，2002）である。

手続き：

まず，レポート原稿に含まれる全ての文に対して，日本語係り受け解析器Cabochaによる係り受け解析を行なう。

例えば，次の例（1）に対するCabochaの解析結果は図1の通りである。

続いて，全ての文に対するCabochaの解析結果に人間が目を通し，Cabochaの解析が誤っていると考えられる箇所を解析し直す。

例（1）に対して人間が解析し直した結果は図2の通りである。

例（1） *スマートフォンなどのパソコンの代わりになるようなものが近年出てきたからである。*

スマートフォンなどの ┐
パソコンの ┐
代わりに ┐
なるような ┐
ものが ──┐
近年 ┤
出てきたからである。

図1：例（1）に対するCabochaの係り受け解析結果

スマートフォンなどの ────┐
パソコンの ┐ │
代わりに ┐ │
なるような ┐
ものが ──┐
近年 ┤
出てきたからである。

図2：例（1）に対する人間の係り受け解析結果

3．結果

日本語係り受け解析器Cabochaによる解析と人間による解析の結果が異なる箇所，すなわち，それぞれが解析した係り先が異なる係り元の文節を，計285件，得た。

それらを「品詞分類」「割込と飛越」「読点の有無」という3つの基準を用いて分類，集計した。

品詞分類：

Cabochaと人間のそれぞれが解析した係り先が異なる係り元の文節について，その文節の最後の形態素を品詞分類した。分類は日本語形態素解析器Mecabを用いて機械的に行い，以下の13種類に分類した。

接続助詞・動詞・助動詞・格助詞・係助詞・
副助詞・助詞副詞化・副詞・名詞・形容詞・
助詞連体化・並立助詞・接続詞

係り元の文節の最後の形態素が動詞または助動詞であるものについては，その係り先の文節の文中での役割が述部に相当するものか格に相当するものかによって区別し，さらに分類した。

その結果，係り元となる文節は次のように分類された。

・述部to述部
［接続助詞・動詞（to述部）・助動詞（to述部）］
・格to述部
［格助詞・係助詞・副助詞・助詞副詞化・副詞・名詞］
・述部to格
［動詞（to格）・助動詞（to格）・形容詞］
・格to格
［助詞連体化・並立助詞］
・接続詞

先述した通り，本研究ではCabocha による解析と人間による解析の結果が異なる場合には，人間の解析結果を正しいもの，係り受け解析器の解析結果を誤ったものとみなして分析を進める。それぞれの分類においてCabocha による解析と人間による解析の結果が異なる場合には，Cabochaによる以下の解析誤りが生じているとみなすことができる。

述部to述部：　複文構造の誤り
格to述部：　格構造の誤り
述部to格：　関係節構造の誤り
格to格：　名詞句構造の誤り
接続詞：　接続構造の誤り

割込と飛越：

Cabochaと人間のそれぞれが解析した係り先が異なる係り元の文節について，Cabochaの解析による係り先，および，人間の解析による係り先のどちらが係り元の文節から近いか，あるいは，遠いかによって分類した。Cabochaが解析した係り先のほうが人間の解析した係り先よりも係り元の文節に近い場合を「割込」，遠い場合を「飛越」と呼ぶことにする。

読点の有無：

Cabochaと人間のそれぞれが解析した係り先が異なるとき，その係り元の文節，Cabochaが解析した係り先の文節，人間が解析した係り先の文節のそれぞれに読点があるか否かによって分類した。

以上の「品詞分類」「割込と飛越」「読点の有無」という基準を用いてCabochaによる解析と人間による解析の結果が異なる箇所285件を分類し，集計したものを表1および表2に示す。

表中，および，以降の本文中では，係り元，Cabochaが解析した係り先，人間が解析した係り先の関係を簡略化して表現するための表記方法として，次のものを用いることにする。

M：　係り元の文節（Modifier）
C：　Cabochaが解析した係り先の文節（Cabocha）

表1： Cabochaと人間の解析結果が異なる頻度（大分類）

	述部to述部	格to述部	述部to格	格to格	接続詞	計
M、H、C。	1	5				6
M、H、C、		1				1
M、H、C＿	1					1
M、H＿C。	8	7		1		16
M、H＿C、	2	2				4
M、H＿C＿	1	7		1		9
M＿H、C。		7				7
M＿H、C、		1				1
M＿H、C＿						
M＿H＿C。	2	11			3	16
M＿H＿C、		4		1		5
M＿H＿C＿	2	11	3	3		19
M、C、H。	1	4				5
M、C＿H。	17	13			2	32
M、C、H、	1					1
M、C＿H、	3	4		1		8
M、C、H＿	1	3				4
M、C＿H＿	7	6		1		14
M＿C、H。	1	3				4
M＿C＿H。	8	27			2	37
M＿C、H、		3				3
M＿C＿H、	1	13	2	1		17
M＿C、H＿		2				2
M＿C＿H＿	6	44	6	17		73
M、H＊C＊	13	22		2		37
M＿H＊C＊	4	34	3	4	3	48
M、C＊H＿	30	30		2	2	64
M＿C＊H＊	16	92	8	18	2	136
M＊＊＊＊＊	63	178	11	26	7	285

表2： Cabochaと人間の解析結果が異なる頻度（小分類）

	述部to述部			格to述部						述部to格			格to格		接続詞	計
	接続助詞	動詞（to述部）	助動詞（to述部）	格助詞	係動詞	副助詞	助詞副詞化	副詞	名詞	動詞(to格)	助動詞(to格)	形容詞	助詞連体化	並立助詞		
M、H、C。			1	2	2				1							6
M、H、C、					1											1
M、H、C＿		1														1
M、H＿C。	1	3	4		2			2	3					1		16
M、H＿C、	2				1				1							4
M、H＿C＿			1	2	1				4					1		9
M＿H、C。					5			2								7
M＿H、C、					1											1
M＿H、C＿																0
M＿H＿C。	2			2	5	1		1	2						3	16
M＿H＿C、				2	2								1			5
M＿H＿C＿	1		1	4	1	2		4			2	1	2	1		19
M、C、H。		1			4											5
M、C＿H。	9	6	2	5	2	1			5						2	32
M、C、H、	1															1
M、C＿H、		1	2	1	2		1							1		8
M、C、H＿		1		2	1											4
M、C＿H＿	3	3	1	3	1		1		1					1		14
M＿C、H。	1				2				1							4
M＿C＿H。	5	2	1	14	7	3		1	2						2	37
M＿C、H、				2	1											3
M＿C＿H、				8	1	1		1	2		2		1			17
M＿C、H＿				1	1											2
M＿C＿H＿	4		2	25	6	3	3	1	6	2	4		9	8		73
M、H＊C＊	3	4	6	4	7			2	9					2		37
M＿H＊C＊	3		1	8	14	3		7	2		2	1	3	1	3	48
M、C＊H＿	13	12	5	11	10	1	2		6					2	2	64
M＿C＊H＊	11	2	3	50	18	7	3	3	11	2	6		10	8	2	136
M＊＊＊＊＊	30	18	15	73	49	11	5	12	28	2	8	1	13	13	7	285

H： 人間が解析した係り先の文節（Human）
。： 句点
，： 読点
＿： 句点・読点なし
＊： 不定（ワイルドカード）
MHCの順序： 文中での文節の並び順

4. 考察

表1および表2の集計結果より，以下のことが明らかになった，

まず，Cabochaによる係り受けの解析誤りのほとんどが割込によるものであり，飛越による誤りは多くはない。割込による係り受けの解析誤りが全体の70%を占め，残りの30%が飛越による係り受けの誤りであった。

・［M＿C＊H＊］（読点のない係り元で生じる割込）
48%（136/285件）
・［M，C＊H＊］（読点のある係り元で生じる割込）
22%（64/285件）

割込による誤りは，Cabochaと人間のどちらの係り先にも読点がない場合に特に生じやすかった。［M＿C＊H＊］（読点のない係り元で生じる割込）の内訳をみると，

・［M＿C＿H。］ 27%（37/136件）
・［M＿C＿H＿］ 54%（73/136件）

だけで81%を占める。また，［M，C＊H＊］（読点のある係り元で生じる割込）の内訳をみると，

・［M，C＿H。］ 50%（32/64件）
・［M，C＿H＿］ 22%（14/64件）

だけで72%を占める。

また，係り受けの解析誤りは，

・格to述部（格構造の誤り）62%（178/285件）
・述部to述部（複文構造の誤り）22%（63/285件）
・格to格（名詞句構造の誤り）9%（26/285件）

の順に多く生じていた。これらだけで全体の9割以上を占める。係り受けの誤りのほとんどは格構造・複文構造･名詞句構造を解析するときに生じたものであり，関係節構造・接続構造の解析において生じる誤りは少ないと言える。

さらに詳細に見ると，係り元に読点がない場合もある場合も，ともに格to述部（格構造の誤り）が多い。一方で，読点がない場合には格to格（名詞句構造の誤り）も，読点がある場合には述部to述部（複文構造の誤り）もそれぞれ多い。

・［M＿C＊H＊］（読点のない係り元で生じる割込）：
格to述部（格構造の誤り）68%（92/136件）
格to格（名詞句構造の誤り）13%（18/136件）

・［M，C＊H＊］（読点のある係り元で生じる割込）：
格to述部（格構造の誤り）47%（30/64件）
述部to述部（複文構造の誤り）47%（30/64件）

以上で確認された傾向を総合すると，係り元・係り先の間に以下のような関係が生じているときに，割込を原因とする係り受けの解析誤りが特に生じやすいことが言える。

［M＿C＿H。］の場合：

読点のない格が文末の述部に係るとき
（格構造の誤り） 27件

「大垣共立銀行が4月11日に$_{M}$「生体認証ATM」を9月下旬に導入する$_{C}$ことを発表した$_{H。}$」

「そして退会しにくくなった後に$_{M}$宗教団体と関連している$_{C}$ことを明かすそうです$_{H。}$」

「オンラインオークションは当初詐欺まがいの取引も多く報告されていたが，出品者の評価制度の導入や第三者の決済を通じて$_{M}$実際に商品が届けられてから$_{C}$決済が起こるサービスなど，いろいろな仕組みが発達したからである$_{H。}$」

［M＿C＿H＿］の場合：

・読点のない格が読点のない文中の述部に係るとき
（格構造の誤り） 44件

「これはグーグルが$_{M}$日本国内を越えて$_{C}$世界中で使用されているインターネットサービスを用いた検索システムの一つであるからだと$_{H}$思う。」

「概要は，スマートフォンが$_{M}$ダウンロードした$_{C}$不正アプリを再生することで$_{H}$，登録された名前や電話番号などがレンタルサーバーに自動的に送信されるしくみだ。」

「そして，近年技術の発達により携帯電話やパソコン

でできることの幅は広がり，大量のデータを持ち運ぶことが可能になっていることを考えると，今まで以上に情報を$_{M}$これらを使う$_{C}$過程で漏洩させてしまう$_{H}$おそれがある。」

・読点のない格が読点のない文中の格に係るとき
（名詞句構造の誤り） 17件

「またツイッターやフェイスブックなどで，エジプトの市民革命や$_{M}$アメリカの$_{C}$若者による格差是正を訴えるためのデモの呼びかけなどが$_{H}$あった。」

「スマートフォンなどの$_{M}$パソコンの$_{C}$代わりになるようなものが$_{H}$近年出てきたからである。」

「しかし，私はツイッターの$_{M}$緊急時における$_{C}$ライフラインとしての活用には$_{H}$賛成である。」

［M，C＿H。］の場合：

・読点のある格が文末の述部に係るとき
（格構造の誤り） 13件

「最後に$_{M}$，これを利用した$_{C}$ことによって起こる不便さだ$_{H}$。」

「この記事を選んだ一番の理由として$_{M}$，私も今年から大学生になり，サークルも何にしようかなと$_{C}$いろいろ考えていたからです$_{H}$。」

「確実に正しいとわかるまで$_{M}$，朝鮮民主主義人民共和国が人工衛星と主張し日米韓などが長距離弾道ミサイルだと推定する飛行物体が発射されたという$_{C}$情報を伝達しようとしなかったのだ$_{H}$。」

・読点のある述部が文末の述部に係るとき
（複文構造の誤り） 17件

「動画や画像などをダウンロードした時，そのダウンロードしたファイルは動画や画像などではなく$_{M}$，個人情報を勝手に流出させるプログラムが仕組まれている$_{C}$ウイルスだ$_{H}$。」

「今後，個人の情報を扱う姿勢は当たり前であるが$_{M}$，情報は危険なものであるという理解をした上で，情報を守ることが必要であると$_{C}$感じたとき，それを託す側と託される側との間に信頼関係ができ，情報はもっと有用で便利なものになるはずである$_{H}$。」

［M，C＿H＿］の場合：

・読点のある格が読点のない文中の述部に係るとき
（格構造の誤り） 6件

「そして「正しい情報」とは，今回のように$_{M}$，それを信ずるに足る確固たる証拠を確認してから$_{C}$発表することだろう$_{H}$。」

「今後，個人の情報を扱う姿勢は当たり前であるが，情報は危険なものであるという理解をした上で$_{M}$，情報を守ることが必要であると$_{C}$感じた$_{H}$とき，それを託す側と託される側との間に信頼関係ができ，情報はもっと有用で便利なものになるはずである。」

・読点のある述部が読点のない文中の述部に係るとき
（複文構造の誤り） 7件

「この事件は情報がどれだけ貴重であるかがわかり$_{M}$，コンピュータに関する知識がある$_{C}$社員なら，会社から情報を取得できてしまうということがわかる$_{H}$事件である。」

「このように進めながら整備に時間がかかるというケースはわかるが$_{M}$，逆に問題が生じる可能性がわかっている上で十分な注意喚起や教育をしないのは企業としての責任を果たしていると$_{C}$いえるのか$_{H}$問題ではないだろうか。」

5. まとめ

最後に，係り受け構造の解析が困難なために文が読みにくくなるような状況を回避するには書き手がどのような対策を取るべきかについて，本研究の結果に基づいて，いくつか提案する。

対策1：
(1a) 係り元の文節が読点のある述部あるいは読点のある格であり，かつ，係り先の文節に読点がないならば，述部として解析されてしまう可能性のある文節を読点なしで割り込ませるべきではない。

(1b) 係り元の文節が読点のない格［格助詞・係助詞・副助詞・助詞副詞化・副詞・名詞］であり，かつ，係り先の文節に読点がないならば，述部として解析されてしまう可能性のある文節を読点なしで割り込ませるべきではない。

(1c) 係り元の文節が読点のない格［助詞連体化・並立助詞］であり，かつ，係り先の文節に読点がないなら

ば，格として解析されてしまう可能性のある文節を読点なしで割り込ませるべきではない。

対策2：
割込が原因で文が読みにくくなる状況を回避するためには，係り元と係り先の距離をできるだけ短く配置するべきである。

対策3：
割込が避けられない場合には，書き手の意図する係り先に読点を配置することで誤りを低減させる必要がある。

参考文献

工藤拓・松本裕治（2002） チャンギングの段階適用による日本語係り受け解析 情報処理学会論文誌，43 (6)，1834-1842.
須藤崇志・丸山広・中村太一（2008） 文を分かりにくくする要因の分析と改善支援手法の提案 電子情報通信学会技術研究報告．SS，ソフトウェアサイエンス，108（64），41-46.
横林博・菅沼明・谷口倫一郎（2004） 係り受けの複雑さの指標に基づく文の書き換え候補の生成と推敲支援への応用 情報処理学会論文誌，45（5），1451-1459.

（2013年9月6日受理）