従属節の階層構造に基づく日本語長文の自動分割とその効果

4L－8　白井　諭[*1]　瀬下 貴加子[*2]　木村 淳子[*2]　横尾 昭男[*1]　池原 悟[*3]

[*1]NTTコミュニケーション科学研究所　[*2]NTTアドバンステクノロジ(株)　[*3]鳥取大学 工学部

1 はじめに

現実の日本語文書には多数の長文が含まれ，構文解析や意味解析などの機械処理を行なう上での障害となっている。特に日英機械翻訳では，長文自体が解析失敗や生成失敗の原因となりやすいだけでなく，英訳されたとしても，原文と訳文の対応が取りにくいため，後編集が困難になるという問題もある。

これらを解決するには長文分割が有効であり[池原 94]，人手による前編集の一環としてしばしば行なわれるほか，前編集支援の課題として検討されてきた[木村 93][松平 94]。しかし，文分割のルールは原文と訳文の比較分析に基づいて経験的に作成されており，日本語原文から見た場合の分割の必然性は不明確であった。

本稿では，従属節の階層的な関係[南 74]に着目することにより日本語の長文を自動的に分割する方法を提案する。具体的には，接続の種類[白井 95b]を考慮することにより適切な分割が可能になること，それにより日英機械翻訳の訳文品質が向上することを示す。

2 従属節の階層構造

日本文においては，書き手の対象に対する認識とその表現過程が，階層的な文構造に反映していることが指摘されている。南は，従属節の述語における助動詞や終助詞の出現情況と従属節の相互関係に着目し，従属節をABCの3種に分類した[南 74]。さらに，①Aは，他のAやBCの一部となれる，②Bは，他のBやCの一部となれるが，Aの一部とはなれない，③Cは，他のCの一部とはなれるが，ABの一部とはなれない，という強い傾向があることを指摘した。しかし，従属節を分類する段階で接続の意味的な関係を考慮する必要があるため，構文解析のような機械処理への適用は困難であった。

これに対し，筆者らは，南の分類の趣旨を生かしながら，意味的判断の困難なものはデフォルトの解釈で分類すること，語尾表現を長単位で分類することなどにより，表1に示すような従属節の再分類を提案した[白井 95b]。この分類によると，Aが2%，Bが92%，Cが6%となり，南の分類に比べるとBの範囲が広く取られる結果となっている。しかし，動作性に着目した再分類と併せて新聞記事文に適用したところ，従属節間の係り受け解析精度は98%となり，十分に精度の高い分類といえる。

以下では，この分類に従って長文分割の方法を検討する。具体的には，表1の連用節のうち，長文分割の対象として適切なものは何かを決定する。

表1 日本語述節の基本分類

形態的分類		機能的分類		備　考
従属節	連用節	A（読点なし）		～ｼつつ，～ｼながら（逆接），～ｽﾙのに続いて
		A（読点あり）		
		B（読点なし）	通常	＜通常＞ ～ｼ，～ｼて，～ので，～ため ＜強中止＞ ～ｼており，～ｽﾙものであり
			強中止	
		B（読点あり）	通常	
			強中止	
		C（読点なし）		～ｽﾙが，～ｽﾙし
		C（読点あり）		
	引用節	引用相当節		～ｽﾙよう(依頼する)
		通常の引用節		～ｽﾙと（発表する）
	連体節	一般名詞型		一般名詞へ係る
		形式名詞型		形式名詞へ係る
主　節		――――		文末の述語句

3 長文分割の着目点

表1の基本分類において，Aは「同時」の表現，Bは「原因」「中止」の表現，Cは「独立」の表現として分類されたものである[白井 95b]。さらに，Bは通常のものと中止性の強いものに分類されている。

直感的には，Cが文分割の対象であると考えられるほか，Bのうち中止性の強いものも文分割の有力な候補である。しかし，新聞記事においては，全述語に対するこれらの述語の出現比率は各6%程度であるため，十分な分割が行えない恐れがある。

ところで，Bには，「原因」に代表される2つの述語間に何らかの論理的関係が見出されるものと，「中

Automatically Splitting Long Japanese Sentences based on Semantically Nesting Clauses

Satoshi SHIRAI[*1], Takako SESHIMO[*2], Junko KIMURA[*2], Akio YOKOO[*1] and Satoru IKEHARA[*3]

[*1]NTT Communication Science Laboratories, [*2]NTT Advanced Technology Corporation and [*3]Faculty of Engineering, Tottori University

止」に代表される話題が展開していくタイプのものとが混在する。「中止」は文分割の候補になりうるが，安易に分割すると文意を不明確にするなどの副作用を生じるため，通常のBについては適用条件を明確にする必要がある。ただし，用言の連用形が単に読点だけを伴う表現（中止形）は，時系列的な複数の事象を表現する際には多用されるので，通常のBのうちでは文分割の第1候補となる。

4 長文分割の効果

本節では，3節で述べた文分割の着目点に対して，実際に長文を分割する実験を行ない，日英翻訳結果に基づいて，長文分割の適否や効果を評価した結果について述べる。

実験に使用した日本文は，日本経済新聞社の有料情報サービスであるテレコンデータベースから抽出した1995年6月の市況速報100記事（計586文，平均35.7文字/文）である。

実際の長文分割処理は，係り受け解析の一部である述語句の認定と分類[白井 95a]を行なうモジュールを流用することにより実現した。そして，日英翻訳システムの内部で，形態素解析結果に対して長文分割処理を適用し，その後，係り受け解析以下の通常の翻訳処理を実行した。文分割の結果，英訳する上で必要な格要素が失われることがあるが，その多くは補完処理[中岩 93]により救済される。

翻訳結果は4段階（秀訳：正しく訳されている，可訳：意味は十分わかる，惜訳：部分的にはわかる，駄訳：まったく理解できない）で評価し，原文訳の評価区分と各分割訳の評価値の違いを集計した。

従属節Cのみを分割対象とした場合の評価値の変化の様子を表2に示す。この分割が行なわれたのは41文（61.8文字/文）で，すべてが2分割であった。表2から，25文字程度の文に分割された場合には訳文品質が向上している。しかし，30文字以上では同等程度にとどまっている。個別に調査すると部分的には改善されているものの，特に40文字以上の文ではさらなる分割が必要であることがわかった。

表2 従属節Cのみを分割対象とした場合
（下段は分割後の文の平均文字数）

前＼後	向上	同等	低下	合計
向上	6文 24.7＼18.2	11文 25.9＼47.6	1文 15.0＼14.0	18文 24.9＼35.9
同等	8文 42.3＼15.3	13文 35.5＼35.2	1文 9.0＼16.0	22文 36.8＼27.0
低下	0文	0文	1文 26.0＼9.0	1文 26.0＼9.0
合計	14文 34.7＼10.0	24文 31.1＼40.9	3文 16.7＼13.0	41文 31.3＼30.5

表3 従属節Bを分割対象とした場合
（従属節Cの分割後，下段は分割後の文の平均文字数）

前＼後	向上	同等	低下	2分割，合計
向上	15文 17.9＼20.1	22文 22.5＼25.0	0文	37文 20.7＼23.1
同等	34文 25.4＼14.8	35文 28.5＼22.6	2文 31.0＼45.0	71文 27.1＼19.5
低下	2文 15.0＼12.0	3文 15.0＼14.7	2文 36.5＼15.0	7文 21.1＼14.0
合計	51文 22.8＼16.3	60文 25.7＼23.1	4文 33.8＼30.0	115文 24.7＼20.3

上上同	上同上	同上上	同同同	3分割，合計
3文 16.7/19.7/24.3	1文 6/16/17	1文 14/12/34	1文 15/12/17	6文 14.2/16.5/23.5

次に，従属節Bのうち，中止性の強いものと中止形のものを分割対象にしてみた。中止性の強いものは11件しか出現しないため，特に分類していない。なお，この分割は従属節Cの分割を行なった後に適用した。評価値の変化の様子を表3に示す。この分割が行なわれたのは121文（45.4文字/文）で，2分割が115文（45.0文字/文），3分割が6文（45.2文字/文）であった。表3から，多くの文が25文字程度に分割されるものの，訳文品質が向上するものは4割程度にとどまっている。しかし，解析処理の成功率はかなり高くなっており，翻訳処理の改良により翻訳精度がさらに向上できる見込みである。

5 おわりに

本稿では，従属節の階層構造に着目して長文を分割する方法を提案し，日英翻訳において訳文品質が向上することを示した。長文分割はいわば前処理的な方法論であり，処理系を改良する方が訳文品質の向上効果は大きい[池原 94]。しかし，長文分割により長文に含まれる複雑な表現の相互影響が極小化されるので，処理系の改良は容易になると考えられる。なお，分割により品質低下が生じたものについては，今後，原因分析を行なう予定である。

参考文献

[池原 94] 池原,白井,小見: 日英機械翻訳における原文前編集の内容と効果について，情報処理学会第49回全国大会 4K-9, pp.3-241-242

[木村 93] 木村,野村,平川: 日英機械翻訳における日本語分割処理について，情報処理学会研究報告 NL-96-8, pp.57-64

[松平 94] 松平: 日英機械翻訳のためのプリエディット支援ツールの開発，情報処理学会研究報告 NL-104-18, pp.135-142

[南 74] 南: 現代日本語の構造，大修館書店

[中岩 93] 中岩,池原: 日英翻訳システムにおける用言意味属性を用いたゼロ代名詞照応解析，情報処理学会論文誌 Vol.34 No.8, pp.1705-1715

[白井 95a] 白井,横尾,木村,小見: 従属節の依存関係を考慮した日本語係り受け解析について，言語処理学会第1回年次大会 A1-3, pp.29-32

[白井95b] 白井,池原,横尾,木村: 階層的認識構造に着目した日本語従属節間の係り受け解析の方法とその精度，情報処理学会論文誌 Vol.36 No.10, pp.2353-2361