超長和文解析支援系

Parsing Support System for Super Long Japanese Sentence

特定非営利活動法人セマンティックコンピューティング研究開発機構理事 **安原　宏**

PROFILE: 2006 年 OKI 定年退職後、ISeC の CDL プロジェクト、Japio の産業日本語プロジェクトに参加

yasuhara@instsec.org

1 はじめに

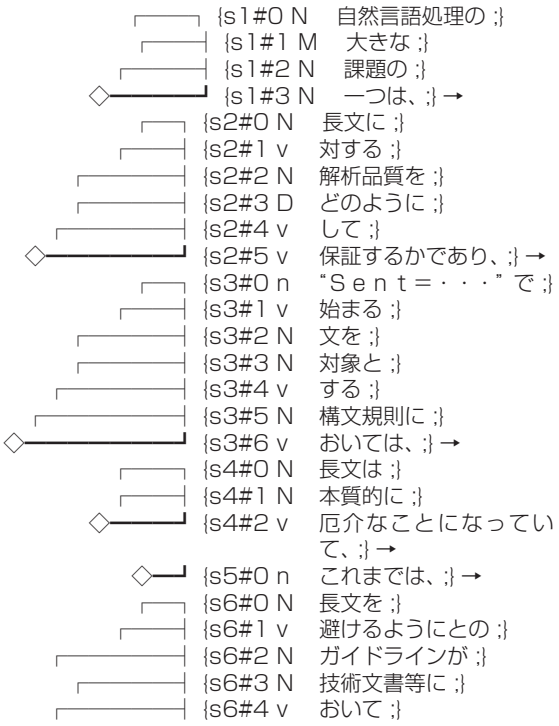
英語や韓国語は分かち書きされているが、IT 時代においても、日本語表記に改善点はないのだろうか。ここでは長文解析の課題を解決しようとする研究とそこから生まれた新しい日本語表記法を示す。

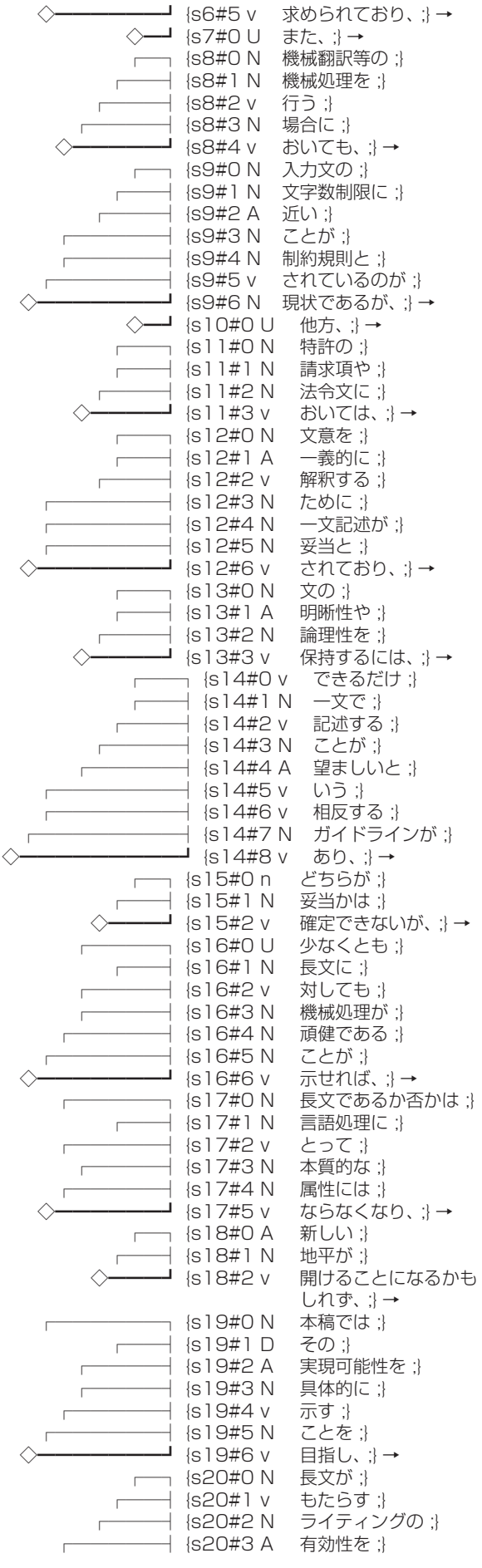
自然言語処理の大きな課題の一つは、長文に対する解析品質をどのようにして保証するかであり、“Ｓｅｎｔ＝・・・”で始まる文を対象とする構文規則においては、長文は本質的に厄介なことになっていて、これまでは、長文を避けるようにとのガイドラインが技術文書等において求められており、また、機械翻訳等の機械処理を行う場合においても、入力文の文字数制限に近いことが制約規則とされているのが現状であるが、他方、特許の請求項や法令文においては、文意を一義的に解釈するために一文記述が妥当とされており、文の明晰性や論理性を保持するには、できるだけ一文で記述することが望ましいという相反するガイドラインがあり、どちらが妥当かは確定できないが、少なくとも長文に対しても機械処理が頑健であることが示せれば、長文であるか否かは言語処理にとって本質的な属性にはならなくなり、新しい地平が開けることになるかもしれず、本稿ではその実現可能性を具体的に示すことを目指し、長文がもたらすライティングの有効性を積極的に活用することを勧め、新しい言語解釈によって日本語が情報処理に置いて他の自然言語と比較してパワフルな言語になりうることを示す。

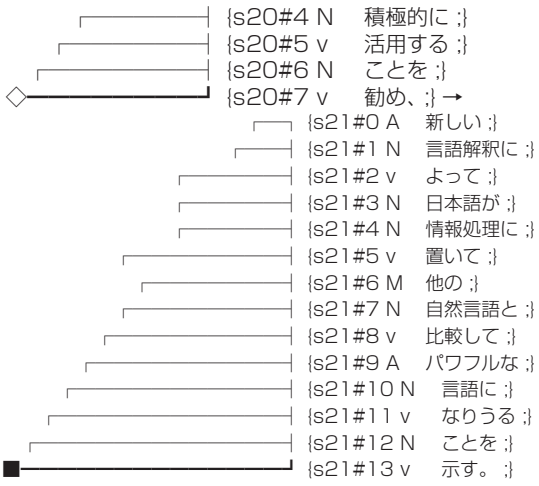
2 超長和文解析

超長和文解析とは、これまでの長文の概念を超越した長さ無制限の和文（超長和文と呼ぶ）を解析することであり、従来から頑健な日本語処理といった表現で、長文に対応する処理系があるが、ここでは、それを極限にまで拡大する解析方法を示す。

上記１の第２段落は、超長和文を意識して記述したものであり、500 字からなる 1 文である。これを解析する処理系は規則ベースでは大半がお手上げである。超長和文解析が与える解は以下のようなものである。







2.1 超長和文仕様

超長和文の言語仕様は、通常の日本語仕様と同じであるが、解析的な観点では、1 文を処理単位とするのではなく、読点を処理単位とすることが特徴である。読点に区切られた範囲を「読点文節」(Punctuation Segment (PS)) と呼ぶ。

例 「超長和文の言語仕様は、通常の日本語仕様と同じである。」の読点文節は、

PS#1：超長和文の言語仕様は、

PS#2：通常の日本語仕様と同じである。

ここで、読点の振り方については自由としている。超長和文解析の蓄積により、読点に関する「正書法」のガイドラインが生まれることを期待している。

2.2 超長和文構文構造

超長和文構文は、読点文節内の文節の係り受け関係（ローカルリンクと呼ぶ）と、読点文節間の係り受け関係（グローバルリンクと呼ぶ）とからなる。リンク情報は、係り先の番号及び係り受け関係である。係り受け関係の最低レベルは、日本語の機能語や接続詞を関係とみなすことができる。上位としては、主格、場所、時間等の格関係、接続関係や論理関係の談話関係等々がある。これらは、対象分野によって適切に設定し、カスタマイズもできる。

ローカルリンクは、読点文節内で閉じたものであり、特徴は、一般に日本語文節の係り受け関係では非交差原理があるが、超長和文の係り受け関係では、これと並んで読点文節からのグローバルリンクは 1 つに限定されるという基本的な原理を課す。つまり、読点文節内の非

末端の文節から外部の読点文節内への係り受け関係がある場合は非文とする。これは、それほど大きな制約でない。本来の読点の用法からして妥当である。この原理は文法の単純さが得られる利点を重視して、採用した。

2.3　超長和文構文の表現

超長和文の構文構造は係り受け（依存）関係であるが、もう一つの課題として、それをどのように表現するかがある。木構造の場合、一般的に 1 文節が 1 行で表現され、その文節の係り先の文節に対してアークがつながるが、係り受けのネストがあると、木構造は横に広がって行き、深い修飾関係が続くときは、見にくくなるという問題点があった。計算機の画面表示を前提にすると、文節が垂直的に並ぶのが読みやすくスクロールも容易だと判断した。さらに、ローカルリンクは文節の左側に配置し、グローバルリンクは文節の右側に配置した。この様な形式が実現できたのは、読点文節を解析単位としたからである。これによって、読点文節内の各文節（自立語と付属語）は完全に垂直に配置され、読点をまたぐ時は、左右にずれる量を極力少ないように設定している。上で例示した超長和文構文表現は、人にとっても読みやすいという効果をもたらす。実際、筆者は、ベタ詰めテキストではなくこの形式でテキストを読むほうが楽に読めることを実感している。

長文構造が解析可能になり、且つ読みやすくなるという点で、「人に理解しやすく、コンピュータにも処理しやすい日本語」を目指している産業日本語とも関係するが、産業日本語の大原則である、「文は短く書きなさい ; Write short sentences」（これは、米国政府の Federal Plain Language Guidelines[1] でも要請されている）は不要になるかもしれない。つまり、短くすると論理関係が曖昧になるという欠点が出てくることが、法令文や請求項が 1 文で記述される根拠とされるが、それをテクニカルライティングにおいても、「文はできるだけ長く論理的に書きなさい」というこれまでとは矛盾するスキルが必要かもしれない。短文か長文かの使い分けは、今後の研究課題である。

2.4　超長和文の解析方法

超長和文解析は、上述したように読点文節を処理単位とし、読点文節内のローカルな係り受け解析を繰り返す。読点文節は文に比較して単語数が少ないので、係り受け構造が単純になり、解析精度が高くなる。本稿のはじめに示した解析例は、京都大学の Juman-6.0 と KNP-2.0[2] を利用した。

読点文節のローカルな係り受け解析ループが終了するとグローバルな係り受け解析を行う。それは読点文節内の末端文節の係り受け解析になるが、デフォルトでは次の読点文節の末端文節を指している。グローバルリンクは、一般的に文と文の接続関係で使用されているもので代用可能と考えている。学習機能を入れることで、ローカルリンク、グローバルリンク共に精度が高くなる。

自動解析が失敗することに対しては、ローカルおよびグローバル解析結果に対して、仮名漢字変換と同様に人手で確認し、係り受けミスを修正できるようにしている。係り先番号の修正や係り受け関係の修正はインタラクティブである。現状は、ローカルリンク及びグローバルリンクの修正は別々に行うが、同時に複数個所の修正が可能である。このようにして100%正確な係り受け解析結果をアプリケーションに提供することができるという点で超長和文解析は自然言語処理における長文問題の解決になる。解析時点では、文節表現は単位要素としているため未知語や文節記載ミスの修正は出来ない。それは、構文解析の前段のテキストエディタに戻って推敲することになる。

3　超長和文構文のさらなる発展に向けて

超長和文構文の効果はより広がりを持っている。メディアという言葉には、映像、音楽とともにテキストも含まれているが、映像メディアと音楽メディアにあって、テキストメディアに欠けているものの一つが「ストリーミング」概念である。テキストストリーミングいう概念は余りポピュラーではない。現状でのテキストメディアは、ベタ詰めのページ概念が主流である。しかし、テキストの表現が紙ベースから電子媒体に移行するにつれてテキストメディアを見直す時期に来ている。そこで、読点文節に基づいたテキストストリーミングの概念を定義してみる。英語等に対しても同様の形式化ができると考

えるがここでは日本語テキストに限定する。

3.1　読点文節ストリーミング

テキストストリーミングの構成要素は、自立語と機能語列からなる「拡張文節」とする。読点文節がブロック（パケット）を形成し、それがあたかも竹の節のように縦にカスケードされてストリームを形成する。表示上は、1 行 1 拡張文節あるいは 1 行 1 読点文節で、後者の場合、拡張文節分かち書きをとる。読点文節の列は、文を形成する。文より上は、通常の分類になる。

例

（読点文節ストリーミング）

{{ 自然言語処理の ;}
{ 大きな ;}
{ 課題の ;}
{ 一つは、;};}
{{ 長文に ;}
{ 対する ;}
{ 解析品質を ;}
{ どのようにして ;}
{ 保証するかであり、;};}
・・・

（読点文節分かち書き）

自然言語処理の 大きな 課題の 一つは、
長文に 対する 解析品質を どのようにして 保証するかであり、
・・・

読点文節ストリーミングファイルを記述する手段としては、ISeC の CDL[3] を用いた。必要なら HTML に変換することも可能である。CDL には係り受け表現のためのリンクプリミティブが存在する点で表現しやすい。

3.2　読点文節オブジェクトモデル

読点文節ストリーミングは、読点文節オブジェクトモデルを提供する。ドキュメントオブジェクトモデル DOM の類似性に着目したもので、自然言語処理の応用展開を加速する API を提供する。

特許にこの API を適用すると、要約書、特許請求の範囲、明細書を含む読点文節ストリーミングが生成されるが、これはベタ詰め特許テキストと比較して、構造化され、形式化されており、リーディング、ライティング、サーチ、トランスレーション等々に統一的なテキストフォーマットを与える。

4　おわりに

文章（テキスト）に読点文節という概念を導入することで、単語数が無制限の長文（超長文）が解析可能になり、この構文構造をベースにして、テキストメディアにストリーミング概念を導入し、そこから種々の自然言語応用がもたらされる可能性を論じたが、特許テキスト処理の新しい方法論になればと考えている。

読点文節や読点文節ストリーミングに関心を持たれた方で、サンプルの実行をご希望の方は、ご連絡いただきたい。

【参考 URL】

[1] Federal Plain Language Guidelines
http://www.plainlanguage.gov/index.cfm
[2] 日本語構文・格解析システム KNP
http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
[3] CDL.core 仕様書 第 1 版
http://www.instsec.org/CDLcoreSpecV1.pdf