文構造解析のための評価観点

富士 秀[1]、長瀬友樹[1]、潮田明[1]、増山顕成[2]

[1]富士通研究所、[2]富士通

fuji.masaru@jp.fujitsu.com

1. 概要

我々の研究では、長文が入力されたときに、文を意味的な構造毎に区切ることを「文構造解析」と呼んでいる。構造解析によって、意味的な単位に区切られた入力文は、人間にとって読みやすくなるのと同時に、機械処理もしやすくなる。
文構造解析システムは、言語処理的な観点から仕様を決めて開発される。このため、システム性能の向上のためには、従来、言語処理的な観点を元にした評価を行いながらチューニングを行ってきた。しかし一方、ユーザの読解性に対してより密接に関連した、読解性評価の観点も重要になってくる。
本研究では、構築した文構造解析システムに対しての、言語処理的な観点および読解性の観点の両面からの評価観点について述べる。また、特許明細書の請求項文を対象としてシステムを実行し、設計した読解性評価観点を用いて評価を行ったのでこれについて述べる。

2. 背景

一般に、文の理解は、入力文が長くなるほど困難になる。言語処理においても、文が長くなるほど係り受けの曖昧性が高くなり、解析精度が低くなる傾向にある。また、人間が読む場合にも、長くなるに従って読解性が下がる傾向にある。
このような問題意識から我々は、入力文を構造部品に区切る、「文構造解析」の研究を行ってきた[1]。
文構造解析では、対象文書が持つ定型性を手掛かりとして、入力文を構造部品に分割することにより、機械が処理しやすく[2][3]人間が読みやすいように加工する。

3. 構築したシステム

これまで研究開発を行ってきた文構造解析技術をもとに、読み手にとっての読解性を向上させるための読解支援システムを構築した。
構築した文構造解析システムでは、さらに、区切られた構造部品を画面上に効果的に配置したり、囲み枠付与や色分け等を施したりすることによって、原文の構成が把握しやすく、各構成要素間の関係が分かりやすくなるようにした。（図 1、図 2）

おもちゃとしての乗物の駆動装置であって、導電性の第1の道路と、前記第1の道路の下に位置した導電性の第2の道路と、第1の道路及び第2の道路と電気的に連絡していて、前記第2の道路上を走行できる動力供給される表面下乗物と、第1の道路及び第2の道路を電気的に付勢することにより前記動力供給される表面下乗物に動力を供給する手段と、前記第1の道路上を走行できる表面上乗物と、前記動力供給される表面下乗物の運動に応答して前記表面上乗物を運動させるよう前記動力供給される表面下乗物と前記表面上乗物を相互に結合する手段とから成ることを特徴とする装置。

図 1.オリジナル入力文の例

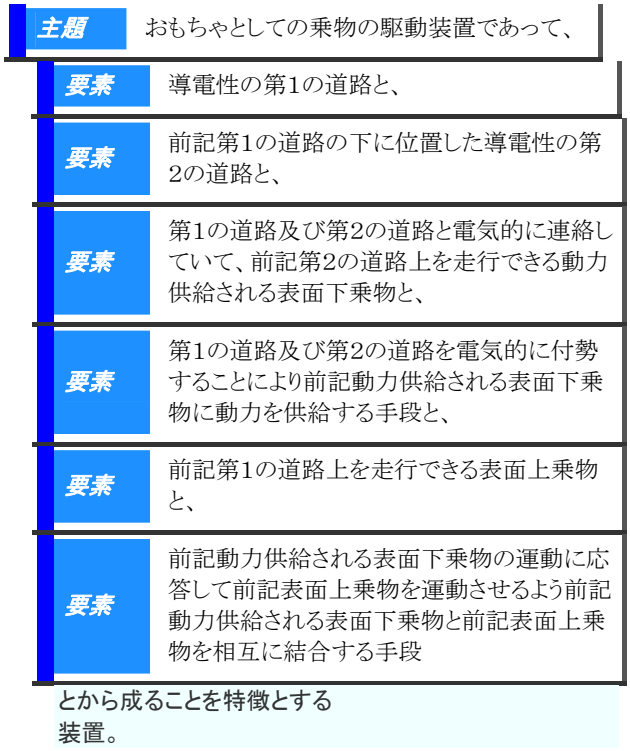

図 2. 文構造解析した入力文の表示例

3.1. 対象文種とシステム構成

本実験の手法は定型性のある文書に対して幅広く適用可能であるが、今回の研究では、日本語特許明細書における請求項（クレーム）を対象として選び、これに対して文構造解析を行うシステムを構築した。

今回の実験では、Perl 言語の正規表現スクリプトで簡便に作成できる範囲でシステム構築を行った。システムは対象文書である特許請求項の定型性に沿って記述されているため、特許請求項に特化したシステムとなった。

3.2. 処理内容

システムでは、入力文である特許請求項は、すべて「主題」、「構成要素」、「説明」の 3 者のいずれかから構成されるものとみなした。（表 1）
システムは、以下の処理を行う。

・**(a)** 入力文を、特許請求項の構造部品である、「主題」、「構成要素」、「説明」に過不足なく切り分けて出力する
・**(b)** 切り出した各構造部品に、その構造部品に対応する「主題」、「構成要素」、「説明」いずれかの構造部品ラベルを付与して出力する

なお、本実験では、論理的な第 1 階層のみを対象とした処理とした。第 2 階層以下の構造は、切り出さずに、つなげたまま出力する仕様とした。（「階層」については、[1]を参照）

表 1. 特許請求項を構成する構造部品

ラベル	構成部品の内容
主題	発明の主題に相当する部分。「○○装置」等の形で表現される。
構成要素	発明を構成する構成要素を表す名詞概念を「構成要素」とした。複数の構成要素が並列で現れることが多いが、スクリプトではこれら並列の構成要素をそれぞれすべて切り出すようにした。
説明	発明の主題および構成要素がどのように動作するかについての記述部分を「説明」とした。説明は、主題を修飾する連体修飾節として現れる場合と、主題を主語としたときの述部として現れる場合の二通りがある。

3.3. 動作フロー

システムの正規表現スクリプト（図 3）は、次のような方針で動作するように記述した。

・文の先頭から順に処理を進める
・文末から文頭方向へのバックトラックは行わない

```
if (
   $line =~ /^(.+?)(と、?)?(を含む|を含んだ|を備える|を備えた|を有する|を有した|を有してなる|を具備する|を具備した)([^、]+?)(であり|であって|においては|において|に於いて|に於て)(、)(.+)$/
   ) {
   $self->{subj_elem} = $1 . $2 . $3;          # 構成要素
   $self->{subj_head2} = $4 . $5 . $6;         # 主部
   $self->{pred} = $7;                         # 残り
  }
```

図 3. 正規表現スクリプトの例

このようなシンプルな構成にしたため、処理性能面では限界もある。つまり、解析に曖昧性があり、文全体のバランスをみながら処理を行わないと正しく切り出せない場合には、不適切な分割がなされる可能性がある。

4. 評価観点

構築した文構造解析システムの評価に先立って、評価観点を策定した。

4.1. 「過分割」と「未分割」

以下の評価観点では、仕様上は本来分割されるべきでない箇所で分割が行われた場合を「過分割」と呼ぶことにする。また、仕様上は本来分割されるべき箇所で分割が行われない場合を「未分割」と呼ぶこととする。

4.2. 言語処理的評価

言語処理的評価は、処理結果と正解の合致度を示す指標であり、システムのチューニングの際に必要となる評価である。

評価対象請求項において、処理結果が完全に仕様通りである場合に、その請求項の構造解析が「成功」と見なすこととした。つまり、過分割も未分割も 1 つもない場合に「○」の判定とすることにした。（表 2）

表 2. 言語処理的評価の判定条件

条件		判定
過分割	未分割	
なし	なし	○
なし	あり	×
あり	なし	
あり	あり	

4.3. 読解性評価

読解性評価は、構造解析前の入力文の読解性を基準として、構造解析の導入による読解性の向上度合いを示す指標である。（表 3）本研究では、構造解析システムを読解性向上のためのツールとして使っているため、その読解性を測定する尺度が必要になる。

表 3. 読解性評価の判定条件

条件		判定
過分割	未分割	
なし	なし	○
なし	あり	△
なし	すべて	—
あり	なし	×
あり	あり	

ここでは、言語処理的評価ではなかった、「△」と「－」の 2 つの条件を新たに導入した。「過分割」は読解性を顕著に劣化させるが、「未分割」は影響が小さいことに着目し、条件を整理しなおしている。

4.3.1. 各判定の例

・「△」の例

過分割は一カ所もなく、未分割はあるものの、少なくとも一カ所で正しい分割が行われている場合に「△」の判定とした。過分割はないため、読解性の劣化は起こっておらず、一箇所以上で読解性が向上している。
図 4.では、「主題」と「説明」が正しく切り出されているために、読解性は向上している。しかし、「請求項 1～3 のいずれかに記載の食物残渣資源リサイクル用システム」において、主題である「食物残渣資源リサイクル用システム」が切り出せていないことから「未分割あり」となったため、「△」の判定とした。

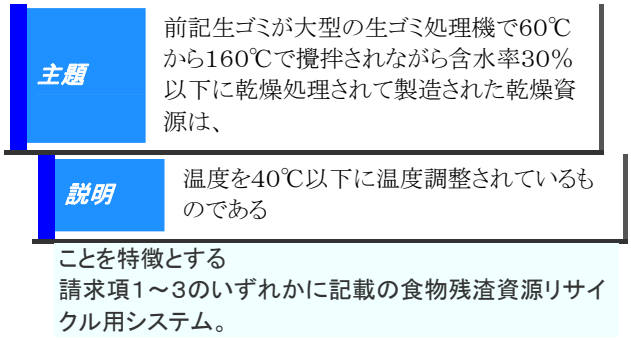

図 4. 「△」の例

・「—」の例

文構造解析前と後とで変化がない場合に、「—」の判定とした。過分割がなく、分割されるべき箇所がすべて未分割の場合に「—」となる。読解性という観点において、文構造解析前と後で読解性に変化がない。
図 5.では、仕様上は、「を接続した請求項 1 に記載の水底土砂の移送装置」において「主題」である「水底土砂の移送装置」が分割されるべきところだが、ここでは未分割となっている。

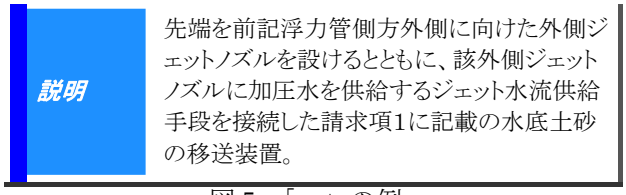

図 5. 「—」の例

・「×」の例

文構造解析結果において、過分割がある場合に、読解性が大きく劣化すると考えられるため、「×」の判定とした。未分割のあり・なしに関わらず、過分割が 1 つでもあれば「×」の判定としている。

図 6.では、「…の方向に対し、」において過分割が起こっている。本来は、「前記切れ目列の方向に対し、」と「その切れ目列を構成する各切れ目の方向が角度を有しており、」の二つの文字列は、つながって一つの「説明」を構成すべきところである。しかし、「…の方向に対し、」が、連用中止形と誤認識され、1 つの独立した「説明」として切り出されている。
これは、本構造解析システムが単純な正規表現に基づいた構成となっているために、「に対し」という前置詞相当語と、「し、」という連用中止形の区別がついていないためである。

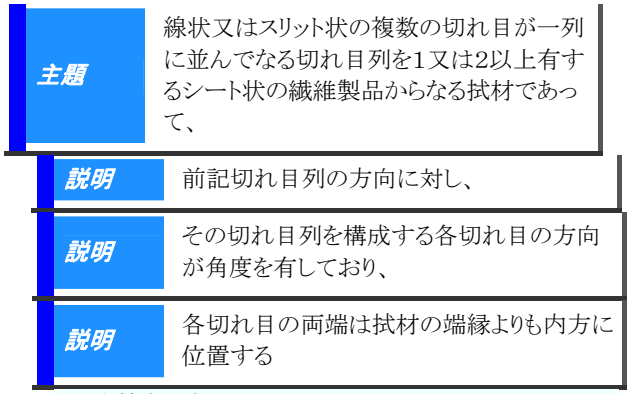

図 6. 「×」の例（分割誤り）

図 7.はもう 1 つの「×」の例である。本来、「主題」と一つ目の「説明」がつながって 1 つの「説明」となるべきところが、「前記第 1 の道路は、」という文字列が誤って切り出されている。
この入力文を正しく構造解析するには、文の全体のバランスをみて、「○○は、…し、」という「説明」の連続で構成されていることを認識する機構が新たに必要となる。

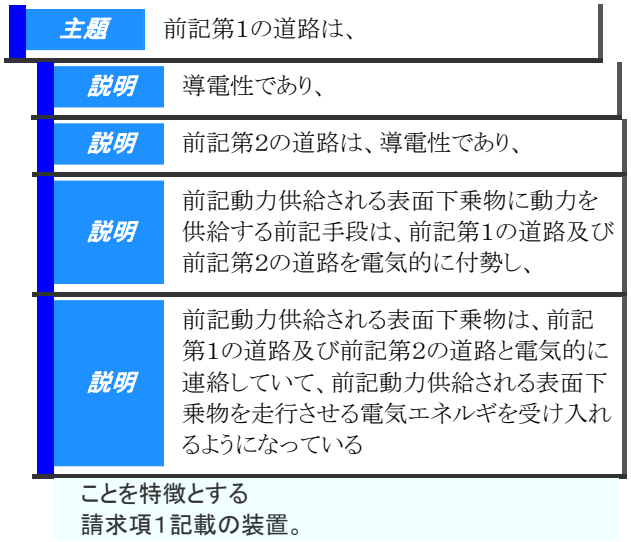

図 7. 「×」の例（全体バランスの問題）

5. 実験方法

5.1. 対象文

本実験では、日本語特許明細書を入手し、そこから請求項を取り出して実験に用いた。各明細書において、最大 10 件の請求項を用いた。請求項が 10 件以内の明細書ではすべての請求項を用い、請求項が 11 件以上の明細書では 10 件で打ち切るようにした。これは、同一の明細書の中では各請求項が均質になる傾向があることから、請求項数が特異的に多い明細書による過大な影響を抑えるためである。

国際特許分類コード（IPC）の先頭文字（A～H）を用いて、実験で用いる対象明細書の分野が均等になるようにした。このように分野が均等になるようにして任意抽出した 200 件の明細書中の請求項を学習データとして用いた。また、また同様にして抽出した別の 200 案件のうち、各案件最大 10 請求項を取り出していって、100 請求項そろったところで、これを評価データとした。

5.2. 事前チューニング

本実験では、日本語特許の請求項を構造解析するためのシステムを用いて実験を行った。実験に先立って、学習データを用いて初期的なチューニングを行った。ある程度のチューニングができた段階で、後述の 1 次評価に入った。

5.3. システム実行

評価データを用いて、1 請求項を 1 入力文としてシステムに入力し、構造解析を行った。各請求項は、取り出した明細書の分野とは関係なく、すべて同列で扱って、実験および評価を行った。

6. 結果

6.1. 1 次結果

システムの初期段階の読解性評価結果を表 4.に示す。

比率的に大きな割合（52%）を占めている「△」の内訳を見てみると、その多くは、ある特定の種類の請求項（従属請求項）に対する正規表現の記述が不十分であることがわかった。また、「×」の内訳を見てみると、システムの構造上の限界に起因するものが多く、修正には大きな手間がかかることがわかった。

表 4. 1 次結果

判定	割合（%）
○	24
△	52
—	14
×	10

6.2. チューニングと再評価結果

1 次評価の結果から、効率よく読解性を向上させるために、従属請求項に対する正規表現の強化に集中してチューニングを進めることとした。また、システムの本質的な問題に関しては、全体の 10%と比較的割合が低いため、今後の課題として後回しにすることにした。

このような方針のもと、再度学習データを用いてシステムのチューニングを行った。この段階を繰り返して、「2 次」および「3 次」の評価結果を得た。

表 5.から、チューニングによって「○」の割合が、1 次の 24%から、3 次の 71%へと、大幅に改善していることがわかる。上記方針に則ってチューニングを行うことによって、読解性を効率よく向上させることができた。

表 5. チューニングによる結果改善

判定	割合（%）		
	1 次	2 次	3 次
○	24	61	71
△	52	17	15
—	14	12	5
×	10	10	9

6.3. 考察

以上説明したように、読解性評価観点を導入したことにより、改善するべきポイントを絞って、これをチューニングによって効率よく改善することができた。また、チューニングした結果、ユーザの読解性がどのように向上するかの見通しを持ちながらチューニングを進められるというメリットがあった。

7. まとめ

人間の読解性向上を目的として、入力文を意味的な単位に区切る「文構造解析」システムを構築した。従来の言語処理的評価に加えて読解性評価の観点を導入し、構築したシステムの評価を行った。この結果、効率的かつ見通しのよいチューニングを実現することができた。

参考文献

[1]富士秀, 長瀬友樹, 潮田明, 増山顕成. 定型性の高い文章に対する日本語構造解析.言語処理学会第 14 回年次大会予稿集,2008.

[2]富士秀, 長瀬友樹, 潮田明, 増山顕成. 原文の定型性を活用した機械翻訳精度向上手法.言語処理学会第 15 回年次大会予稿集,2009.

[3]富士秀, 長瀬友樹, 潮田明, 増山顕成. 部品化された原文からの機械翻訳文生成.言語処理学会第 16 回年次大会予稿集,2010.