定型性の高い文章に対する日本語構造解析

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

定型性の高い日本語文章の構造的な特徴を解析することによって精度良く文章構造解析を行うシステムを構築しその有効性を検証した。産業分野で扱われる文書では分野独特の定型的な表現が存在することが多いが、この定型性を利用することによって構造解析精度が向上することが期待される。
本研究では、表層的な定型表現を処理するための広域的パターン処理と、文節・係り受け等の局所的な解析処理とを統合することによって、定型性の高い文章を精度よく構造認識するための枠組みを作成した。構築したシステムは、対象分野にチューニング可能なルール記述の枠組みを備え、入力文を構造部品に分割し各構造部品にラベルを付加して出力する。

2. 背景

一般的に現状の言語処理技術では、短文に対する解析精度はそれなりの実用レベルに達しているものの、長文に対する精度は不十分である。

本研究が対象とする産業分野で扱われる文書（例えば、特許文書や契約書、等）は、長文で構成されることが多いが、記述された内容を観察してみると、分野固有の定型的な表現が多用されることがわかる。そこで、この定型的な特徴表現を手掛かりとして入力文を適当な長さの意味的な単位に区切り、その単位に対して言語処理を適用することによって処理精度の向上が期待される。

言語処理技術の例としては、翻訳支援[1][2]における機械翻訳がある。翻訳支援の利用場面では、入力文を適当な短い単位に区切り、その単位に対して機械翻訳を適用すると、実用レベルの翻訳精度が得られる確率が大幅にあがると考えられる。

3. 目的

本研究の目的は、入力文の構造的な特徴を解析することによって、入力文を**「構造部品」**に分割し、各構造部品に対してその役割を表す**「ラベル」**を付与する技術を開発することである。

特許要約文を例にとって、本システムが目標とする出力例を図 1. に図解する。「主題」の構造部品では、「であって、」という表層的な特徴表現を区切として、「翻訳支援装置」という主題を切り出している。「主題」に係る「説明」の構造部品は、「行う」という連体動詞を特徴表現として、切り出しを行っている。

ラベル	構造部品
説明	第1言語での入力文に基づいて、第1言語の例文とその例文の第2言語による翻訳文を対訳例文として検索し翻訳支援を行***う***
主題	翻訳支援装置***であって、***
要素	入力文を受付ける入力文受付部***と、***
要素	受付けられた例文の部分列を作成する部分列作成部***と、***
要素	例文の部分列を用いて第1言語と第2言語の対訳例文を検索する対訳例文検索部***と、***
要素	対訳例文と該対訳例文に係る例文の部分列とに基づいて、検索された前記対訳例文の評価を行い評価値を付与する評価値付与部***と、***
要素	評価値に基づいて、対訳例文から所定の対訳例文をフレーズ候補として抽出するフレーズ候補抽出部***と、***
要素	フレーズ候補から所定のフレーズを選択するフレーズ候補整理部***と、***
要素	フレーズ候補が付加された前記入力文を表示するフレーズ候補付き入力文表示部***とを***
要素動詞	備える。

図 1. システム出力例（特徴表現を下線で示す）

4. 課題

本研究が対象とする文章では、その分野におけるある種の書き方の慣習が存在し、書き手もその慣習に則って文書を作成する。分野によっては、書き方の手引書が存在する場合もある。その結果、作成された文章は、人間が読む範囲では、ほぼ完全な定型性があるかのように見える。

しかしながら、制限言語エディターを用いる等の特殊事情の場合を除いて、書き手は文章を自然言語で自由に記述するため、作成された文章は、機械処理の観点では曖昧性が存在することになる。

この曖昧性のために、特徴表現を表層的に捉えるだけの単純なパターンマッチングでは枠組みとして不十分である。本研究では、以下の 2 点の曖昧性に着目し、それらを解決するためのシステムを設計した。

階層構造による曖昧性

曖昧性の多くは、文章中に複数階層の構造が存在する場合に発生する。単純なパターン処理だけでは、定型性のきっかけとなる表現が、どの階層に属するものであるかを判定することができない。

例えば、「○○であって、」という特徴表現によって主題の区切りを表す文章があったときに、「○○」が文全体の主題であるのか、もしくは文中の１つの節の主題であるのかは、文全体のバランスを考えなければ判別できない。

特徴表現の揺れ

文章作成で用いられる特徴表現は、ある程度固定であるものの、自然言語で記述されるがゆえの揺れが生じる。単純な表記のマッチングだけではこれら表記の揺れに対応することはできない。
また、表記に変化が起こるような特徴表現に対応できる必要がある。例えば、連用中止を特徴表として用いたい場合、実際の表記は「～し、」であったり、「～り、」であったりと、手がかり表記は異なることになる。

5. 解決手段

上記課題を解決するために、文全体の構造パターンを考慮しながら複数階層を扱う仕組みと、言語処理による揺れを吸収する仕組みを、同時に実現する枠組みを考案した。
なお、この枠組みは、任意の複数階層に拡張することも可能だが、今回の実験では、**第 1 階層のみを分割対象として**実験を行った。第 2 階層以下の構造は、特徴表現において分割されずに出力される。
ここで「第 1 階層」とは、文全体の主題に直接関係する構造のことであり、「第 2 階層」は第 1 階層への埋め込み文、「第 3 階層」は第 2 階層への埋め込み文、…である。

複数階層への対応

対象文中には複数の階層の特徴表現が混在するが、対象階層（ここでは、第 1 階層）の構造を正しく把握するためには、対象文全体の構造パターンをみて、どの特徴表現が対象階層の特徴表現であるかを判断する必要がある。

この課題を解決するために、最初に、どの階層に属するかは度外視して、表層的に拾える特徴表現をすべて拾い出しておく。これら特徴表現で切り出された単位を、以降**「小節」**と呼ぶことにする。そして、文全体の構造パターンと照合しながら、どの小節が、第 1 階層に属すものであるかを判定するような枠組みとした。

特徴表現の揺れへの対応

特徴表現とのマッチングに先立って、入力文中の様々な表記揺れを処理しておく必要がある。ここでは従来の文節処理技術を用いて、入力文を、正規化された文節列に分解する。一般的に文節レベルの処理精度は高いため、文節を処理の基点とすることは有効であると考えられる。
なお、文節には、文法属性を付与しておく。これによって、例に示した「連体動詞」や「連用中止」のような特徴表現を記述できるようになる。

6. システムの構築

以上のアイデアを実現する日本語構造解析システムを構築した。システムは、入力文章を構造部品に分解し、分解した各構造部品にラベルを付与して出力するものである。

6.1. 中間形式

構築したシステムでは、従来型の自然言語解析の結果である「文節」を最小単位とし、最終的な構造解析の結果である「構造部品」の間に、「小節」という単位を設ける。この「小節」が、局所的な言語処理結果と大域的な構造パターン処理の橋渡しをする。

文節

従来の日本語解析で得られる、いわゆる「文節」である。文法属性が付与されている。

小節

複数の文節が組み合わさって「小節」を形成する。最終的な構造部品の区切り箇所となるかどうかとはまずは無関係に、文節列を、出現する全ての特徴表現の出現位置で区切ったものが小節である。

構造部品

構造部品は、全体システムの最終出力結果である。複数の「小節」が組み合わさって、「構造部品」を形成する。文全体の構造パターンとの照合の結果、第 1 階層に対応する小節のみが生き残り、第 2 階層以下の複数小節が結合されて、最終的な構造部品を形成する。

6.2. ユーザ定義ファイル

前項で説明した中間形式によって実際に表される内容は、対象分野毎に異なったものとなる。そこで本システムでは、ユーザが対象分野に合わせた定義ファイルを作成することにより、システムを外部から自由にカスタマイズできるようにした。

小節定義ファイル

小節は特徴表現によって区切られるが、特徴表現は分野に依存したものである。ユーザは、対象分野の特徴表現を記述した小節定義ファイルを作成することによって、分野に対するカスタマイズを行うことができる。システムは、実行に先立って、小節の区切りとなる特徴表現を記述した「小節定義ファイル」を読み込む。小節定義ファイルは、システム利用者にとって可読な記述形式で記述できるようになっており、プログラム実行に先立って、ローダによって、システムが実行可能な内部規則に変換される。

構造パターン定義ファイル

構造パターンは、文中の構造部品の並びのパターンであり、これもまた分野に依存したものである。前項の小節定義ファイルと同様に、構造パターン定義ファイルも、可読な記述形式になっており、ローダを通じてシステムに読み込まれ、実行される。

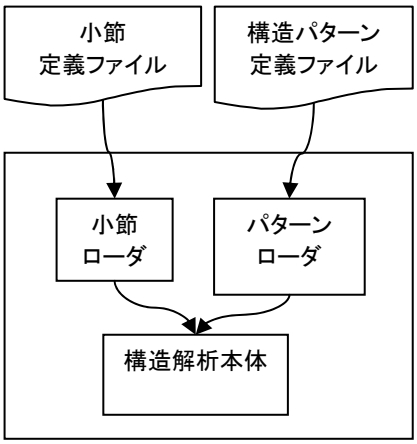

図 2. システムと定義ファイルの関係

6.3. システム内部の構成

以下は、前述の定義ファイルを読み込んだ状態で、システムが最終的に構造部品列を出力するのに必要なシステム内部構成である。

文節生成

入力文に対する文節列を生成する。各文節には、後段の小節処理で利用できるように、文法属性が付与されている。

小節生成

ローダによって内部形式に変換された小節定義を参照しながら、文節列を、小節列に組み上げる。特徴表現が小節の右端にくるように、文節をつなげていく。小節生成では、特徴表現の表記だけではなく、文節の文法属性も参照できるようになっている。

構造候補生成

ローダによって内部形式に変換された構造パターン定義を参照しながら、小節列を、構造部品列に組み上げる。このとき、複数の構造部品列の候補が生成されるが、これらが、複数構造候補となる。

構造候補選別

作成された構造候補について、各構造部品の妥当性をチェックし、妥当でない構造部品をもつ候補を構造候補から除外する。各構造部品は、複数の小節から構成されているが、これら複数小節間の係り受けの妥当性をチェックする。これによって、文法的に妥当な構造部品をもった候補のみが残ることになる。

構造候補ランキング

最終的に残った構造候補対して、構造の確からしさに応じた評価値を付与し、評価値によってランキングしてからユーザに構造候補を提示する。

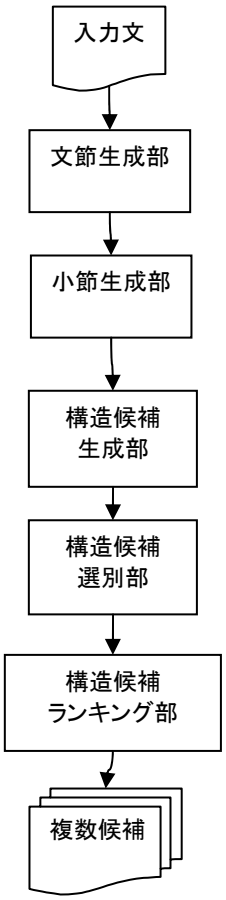

図 3. システムの内部構成

6.4. 構造候補のランキング

構造候補選別部によって選別されて残った候補は、構造部品の組合せとしても妥当で、かつ各構造部品の中の構文的な整合性も妥当である。これら候補の中で、意味を斟酌しながら人間が読んで正解として認識するものは、一般的に 1 つに絞られる。（ただし人間が読んでも 1 つに絞れないものも少数ある。）
本来は、文章の意味にまで立ち入らなければ正解は得られないが、全般的な傾向としては、構造部品のバランスのよい候補が、正解候補である可能性が高い。そこで、構造部品のバランスを考慮した評価関数を用意し、各候補に付与された評価値によって候補のランキングを行うようにした。

構造部品のバランス評価で最も効果の大きいものが、並列構造のバランスの良し悪しである。例えば、並列関係にある構造部品がすべて等しい右端表記を持っている場合に、高い評価値を与えるような関数を用意した。

7. 評価実験

開発した構造解析システムに対して構造解析精度を測定するための実験を行った。

評価対象文書

特許庁の特許明細書検索ページから、弊社を出願人とする自然言語処理関係の発明を無作為に 50 件抽出し、ここから「出願人要約」を取り出して評価例文とした。要約文は、「発明の名称」、「課題」、「解決手段」から構成されるが、それぞれを切り出して、別々に評価することとした。このうち「発明の名称」と「課題」は必ず 1 文から構成されるが、「解決手段」は 1 文の場合と複数文の場合がある。

定義ファイルの準備

実験に先立って、対象分野に即した、小節定義ファイルおよび構造パターン定義ファイルを作成した。オープン評価となるように、評価対象の発明を参照することなく作成作業を行った。

システム実行

作成した定義ファイルを組み込んだ状態で評価例文を入力し、出力結果を得た。1 つの入力について、最大 10 候補を出力するようにした。

評価

各入力文について、評価者の考える構造部品・ラベルと、システムが出力した構造部品・ラベルが一致した場合に正解とした。各入力について、候補中に正解が存在すれば「○」、しなければ「×」という基準で評価を行った。人間が読んでも正解が一意に決められない場合は、複数の可能性のいずれかと一致していれば「○」とした。

8. 結果と分析

表 1 に、特許要約 50 件に対する構造解析の正解率を示す。出力された 10 候補における正解率を、入力文タイプ別に表にしてある。

表 1. 構造解析の正解率

	正解率		
	発明の名称	課題	解決手段
1 位	88%	24%	64%
3 位以内	98%	58%	86%
10 位以内	98%	70%	90%

9. 考察

9.1. 入力文タイプ別の分析

「発明の名称」、「課題」、「解決手段」のうち、「発明の名称」は、基本的には名詞句であって文自体が短く、高い精度で解析ができている。「課題」も 1 文で構成されるが、意味的な曖昧性が大きいため、精度が低目となっている。「解決手段」は、3 者の中でもっと文が長く、人間にとっては扱いにくいが、構造の定型性が比較的高いため、比較的高い精度が得られている。

9.2. 精度向上の可能性

構築したシステムは、本研究の課題を概ねクリアするものとなった。本質的な枠組みの問題は現時点では認められておらず、以下の点を改良することにより、さらなる精度向上が見込まれる。

構造候補選別部のチューニング

構造部品内の構文チェックは、構造部品内の小節間の係り受け可否によって行っている。しかし、文節生成のために用いた従来型の解析ツールは、今回のような小節単位の処理を想定していないため、係り受け解析に必要な属性が適切に付与されていない場合がある。本システムの用途に合わせたチューニングがある程度必要になる。

構造候補ランキング機能の増強

実験結果では、10 位以内で見ると正解率はそれなりに高いが、実用面を考えると、もっと上位に正解を集めたい。このためには、文全体のバランスを評価するための関数を充実させ、ランキング精度を向上させる必要がある。

10. まとめと今後

長文の入力文を、分野固有の特徴表現を区切とした構造部品に分解する日本語構造解析システムを構築し、評価実験を行った。その結果、実用面において十分な品質の構造解析結果を得ることができた。システムの改良を進めてさらなる精度向上を図りたい。また、システムの分野拡張性についての検討を進めたい。

なお、本システムの結果は、後段の言語処理システムへの入力として使われるが、後段のシステムとも合わせたトータルな評価も実施していく。

参考文献

[1]潮田明,富士秀,大倉清司,山下達雄.機械翻訳と訳例検索を統合した翻訳支援システム.言語処理学会第 9 回年次大会予稿集,2003.

[2] 長瀬友樹、大倉清司、富士秀、徐国偉、潮田明. 多言語翻訳プラットフォームによる翻訳サービスの実装. 言語処理学会第 12 回年次大会予稿集,2006.