

6 L - 4

日本語文構造分類ツール

山端 潔 赤峯 享 奥村 明俊

日本電気(株) C & C 情報研究所

1. はじめに

自然言語処理システム開発は、言語データを収集・分類・解析し、その結果に基づいてシステムを改良し、評価し、明らかになった問題点について再び言語データの収集を行う、といった一連の流れの繰り返しで行われる。それゆえ、言語データの収集・分類・管理は、言語現象解析のための原資料収集としての、あるいはシステムの性能評価のための基準作成として大きな重要性を持つ。

現象の網羅性、さらにはその出現頻度分析をも要求しようとすれば、データとして膨大な量を収集することが必要である。従ってデータ整備作業はできるだけ機械化することが望まれる¹⁾。本稿では自然言語研究開発支援環境²⁾の一部として、日本語の統語現象別分類を行うために構築した言語データ加工・構造分類支援ツールについて述べる。

2. 機能概要

性格と機能を定めるためにはその使用目的を明確にしておくことが必要である。本稿のツールの設計においては以下の二点を主要な目的とする。

a) 日本語の統語的振舞いを定量的に研究するための支援ツールとして使用する。特に、日本語の構文規則を開発するための基礎的な言語データを作成するためのツールとして使用する。構文規則の開発のためには、対象言語の言語現象を網羅的に収集し、加工して検索しやすい形でまとめておく必要がある。加工としては、最低限形態素解析情報の付与が必要である。検索は種々の形で行える必要があるが、おおきくは統語的な振舞いの別に分類するのが便利である。開発しようとする解析規則自体が統語的振舞いと同じくする現象毎にまとまっているからである。

b) 自然言語処理システムの処理能力を検証するための日本語文例集を作成するために使用する。システムを統語現象別に検証するために文を加工・分類し、例文を蓄積してゆく環境を構築するために用いる。

以上のような目的を達成するために次のような分類基準を設け実現した。

2. 1. 分類機能の概略

第1の分類軸は、文中の用言数とその性質である。具体的には

- 1-a) 文中に用言がいくつ存在するか。
- 1-b) 各用言は名詞を修飾しているか否か。
で第1の分類を行う。軸a) はおよそ文の複雑さに対応する。軸b) は、名詞修飾の用言が出現すると格

要素及び格関係の認定に困難が大きく増すことを考えて設けられた軸である。これらは、全体として英語で「單文・重文・複文」といわれる分類にほぼ対応しているが、引用文等は本稿では埋め込みと見なさない点がそれらとは異なる。

入力文はまずこの軸1を基準に該当する部分に分割され、大分類されたのち、次の分類軸2を代表とする分類軸によりさらに加工・細分類される。

第2の分類軸は、主文節(用言、体言)に対する機能語の種類と出現順序である。たとえば、

「私が彼女に会う。」
は主用言として「会う」をもち、それに対する機能語は「が」と「に」である。あるいは、名詞句
「人民の人民による人民のための政治」
では主用言は「政治」であり、それに対する機能語は「の」「による」「のための」である(ただし、「による」と「のための」を助詞相当語とした場合)。

ここに述べたような機能を実現するために、辞書と形態素解析ツールとして用い、各文節にふられた統語情報を参照しながら分類を進める。

使用した情報としては次のようなものがある。

- 品詞及びその細分類
- 格助詞、副助詞、接続助詞の種類等、付属語の情報
- 用言の活用および必須格情報
- 表層語彙

2. 2. 分類機能の詳細

以下は軸1にそった具体的な切り分け基準の例である。

- a) 名詞句
文中に用言をまったく持たないもの。ただし、格要素を一つしか持たない形容詞・形容動詞の連体形は用言と数えない(この基準は以下同じ)。
- b) 単文
文中に用言を一つだけ含み、それが名詞を修飾していないもの。
- c) 埋め込み名詞句
文中に連体形の用言を一つだけ含む名詞句。
- d) 埋め込み単文
文中に連体形の用言を一つだけ含む単文。
- e) 重文
文中に複数の用言を含むが、そのどれも名詞句を修飾していないもの。
- f) 複文

上記以外の全ての文。

これらに相当する部分を一文中から切り分けるために、文中の部分の用言の数と種類を調べる。ただし、

1. 形容詞のように必須格の数が一つであり、連体形のものは用言として数えない
2. 「名詞+だ」のものは用言として数える。

は例外とする。

分類の際に問題となるのは埋め込み区間の開始位置をどう同定するかだが、これは、埋め込み用言の必須格の数、助詞の「は」と「が」の使い分け、読点の位置等のヒューリスティックスによる推定を行う。

この方法を用いて実際に切り分けを行った結果を以下に示す。

入力文) 「私は彼女の兄が買った本をもらったが、その本を捨てた。」

a) 名詞句：

「私は」、「彼女の兄が」、「本を」、
「その本を」

b) 単文

「私は本をもらったが、」
「その本を捨てた。」

c) 埋め込み名詞句

「彼女の兄が買った本を」

d) 埋め込み単文

「私は彼女の兄が買った本をもらったが、」

e) 重文

「私は本をもらったが、その本を捨てた。」

f) 複文

「私は彼女の兄が買った本をもらったが、その
本を捨てた。」

なお、この先の処理では文末の不用な助詞等の情報は捨てる。

次に第2軸にそった切り分け基準の例を挙げる。

a) 名詞句

- 機能語の種類と出現順序による分類

第一段階で切り出された名詞句中、その品詞が付属語として登録された語を抽出し、その種類と出現する順序で分類し、インデックスファイルを作成する。

- 複合名詞句の構成単位である単位名詞句の種類と出現順序による分類。

複合名詞句を構成する単位名詞がもつ名詞としてのカテゴリーの細分類をキーにその出現順序でインデックスファイルをつくる。

b) 単文

- 機能語の種類と出現順序による分類

各格要素にそれを持つ格助詞、副助詞、並立助詞を対応させ、それらの文中での出現順序をキーにして文を分類し、インデックスファイルを作成。

- 不用修飾の除去を行った文を出力

文中の名詞句への修飾語、たとえば最初に用言でないと認定された形容(動)詞や副詞および名詞句構成要素のなかで主要でないものを削除。

c) 埋め込み名詞句

- 機能語の種類と出現順序による分類

埋め込み主用言に対する機能語を単文の場合と同様に抽出してインデックスファイルを作成

- 不用修飾の除去を行った文を出力

d) 埋め込み単文

- 機能語の種類と出現順序による分類

主用言に対する機能語を抽出して分類。

- 不用修飾の除去を行った文を出力

e) 重文

- 用言数による分類

用言数でカウント

- 連用中止形または連用中止形+「て」の連続パターンの認識

- 不用修飾の除去を行った文を出力

2. 3. ユーザーインターフェース

このような言語ツールでは、ユーザーインターフェースも重要である。特に、言語現象解析の補助ツールとして使用する際には大量のデータの間を自由に行き来できることが必要である。このためにXウィンドウ上に表示機能を主とするユーザーインターフェースを構築した。

3. 終わりに

本稿のツールはその初期目的を達成するための最低不可欠な機能を提供しているが、更に改良を加える必要がある点には次のようなものがある。

- 切り分け、とくに連体節の同定の高精度化のためのヒューリスティックスの改善
- ユーザーインターフェースの高度化

これらについてはツールを使用しながら徐々に改善を図ってゆく予定である。

謝辞

ツールの開発を担当していただいた日本電気技術情報システム(株)の浜田和彦氏に感謝の意を表す。

参考文献

[1] 小倉他：「言語データベース収集支援システム」、情処第36回全国大会(1988)

[2] 飯野他：「自然言語研究開発支援システム」情処第39回全国大会(1989)