

日本語解析用クラスライブラリ「Jpn クラスライブラリ」の開発

4Q-5

小松順子

(株)リコー 情報通信研究所

1 はじめに

日本語解析はテキスト音声合成を始め様々なアプリケーションに応用できる。そこで、日本語解析系を実装した汎用的なソフトウェアライブラリの存在が望まれる。

我々は、従来から汎用的な日本語解析用ライブラリを開発を試みてきたが、従来のソフトには次のような問題点があった。アプリケーション固有の処理との融合がうまくいかず、アプリケーションソフトの処理効率やメンテナンス性が低下してしまう。ソフトウェアモジュールの切り口が少ないために、アプリケーションのニーズに合わせて、日本語解析に関連する様々な処理を組み合わせることが困難で、アプリケーション開発の幅を狭めてしまう。

そこで、日本語解析系をオブジェクト指向に基づいてモデル化し、クラスライブラリ「Jpn クラスライブラリ (Jpncl)」として実現した。継承を使うことによって、アプリケーション固有の処理との融合が容易になり汎用性が高い。また、クラス化したことによって、モジュールの切り口が増え、再利用性が高いので、アプリケーション開発の幅の広がりも期待できる。本稿では、Jpnclの基本となる日本語解析系のモデル化を中心に述べる。

2 日本語解析系のモデル化

ここで日本語解析系とは、形態素解析、係り受け解析とそれらに関連する言語データ群（単語辞書、文法規則など）を指す。また、形態素解析、係り受け解析の基本アルゴリズムは逐次確定型のコスト最小法に基づいている。[1]

このような日本語解析系を言語プリミティブ、

A Development of the Class Library for Japanese Text Analysis, "Jpn Class Library"
Junko Komatsu

Information and Communication R&D Center
RICOH Co., Ltd.

言語解析系、言語データ、日本語文字関連、品詞関連の5つのクラス群でモデル化した。言語プリミティブクラス、言語解析系クラスについては、次章以降で説明する。言語データクラスには、単語辞書クラスや接続表クラスなどがある。日本語文字関連クラスには、日本語文字クラス、日本語文字列クラスがあり、SJIS、EUCなどの漢字コード体系の違いを吸収している。品詞関連クラスには、品詞テーブルクラス、品詞クラスがあり、階層的な品詞分類をサポートしている。

2.1 言語プリミティブクラス

言語プリミティブクラスを図1に示す。以下の図はOMT法[2]の表記法に準ずる。言語プリミティブクラスには、コスト値を持つ解析候補を表す抽象クラスJpnObjectを設け、その下に単語クラス(Word)、文節クラス(Bunsetsu)、句クラス(Phrase)を設けた。

単語クラスには、形態素解析に最低限必要な情報のみが含まれ、解析結果に対して付加したい情報は、単語定義クラス(WordDef)に持たせる。文節クラスは文節を構成する単語オブジェクトの列を持ち、句クラスは句を構成する文節オブジェクトの列を持つ。

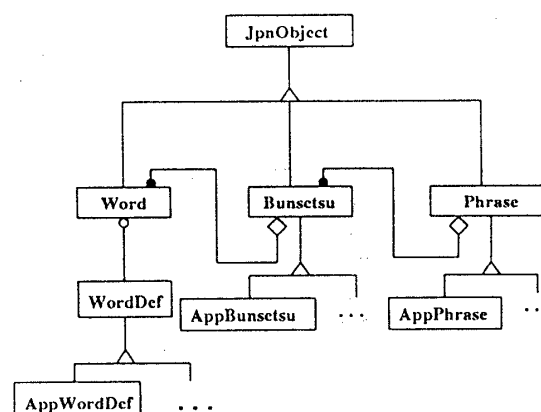


図1: 言語プリミティブクラス

アプリケーション固有の情報や処理は、単語定義クラス、文節クラス、句クラスを継承させてアプリケーション専用クラス（AppWordDef、AppBunsetsu、AppPhrase）を作成し、その中に実装する。これによって、アプリケーション固有の処理を無理なく融合することができる。

2.2 言語解析系クラス

言語解析系クラスには、コスト最小法に基づいて解析候補の絞りこみを行うラティスクラス（JpnLattice）（図2）と、解析候補を生成するジェネレータクラス（JpnGen）（図3）がある。

ラティスクラスの下には、単語候補間の接続コストを接続表（CnctMtx）から求めて、単語候補を絞りこむ単語ラティスクラス（WordLattice）、係り受け規則（KuRule）や結合価辞書（KframeDict）を基に文節候補間の係り受けコストを求めて、文節候補を絞りこむ文節ラティスクラス（BsLattice）を設けた。

ジェネレータクラスは、解析候補となる言語プリミティブオブジェクトを、それより小さい言語プリミティブオブジェクトから生成する機能を持つ抽象クラスである。その下には、単語オブジェクトを生成する単語ジェネレータクラス（WordGen）、文節オブジェクトを生成する文節ジェネレータクラス（BsGen）、句オブジェクトを生成する句ジェネレータクラス（PhraseGen）がある。

文節ジェネレータを例にして、ジェネレータクラスの機能を説明する。文節ジェネレータは、単語ラティスと単語ジェネレータへの関連を持っている。そして、単語候補を受け取るとそれを単語ラティスに格納する。一方、単語より小さいオブジェクトを受け取ると、それを単語ジェネレータに渡し、単語候補を生成させる。そして、最終的に、単語ラティスからもっともらしい文節候補を取り出して出力する。

このような構成にすることによって、文字（または文字列）、単語候補、文節候補が混在した入力を扱える日本語解析系を実現することができる。

3 クラス設計およびインプリメント

クラス設計では、前述のモデルで示したクラスの他に、実装のしやすさを考慮して、いくつかのクラスを追加し、約50のクラスを設定した。イン

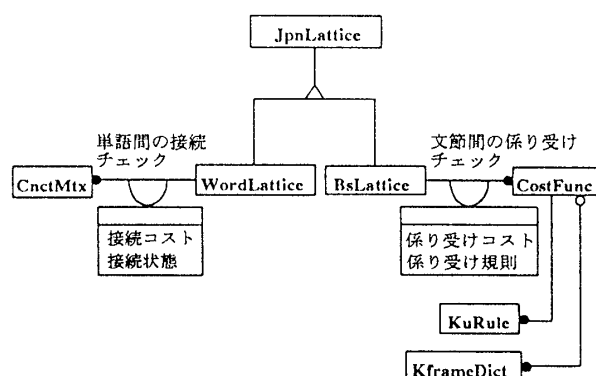


図2: ラティスクラス

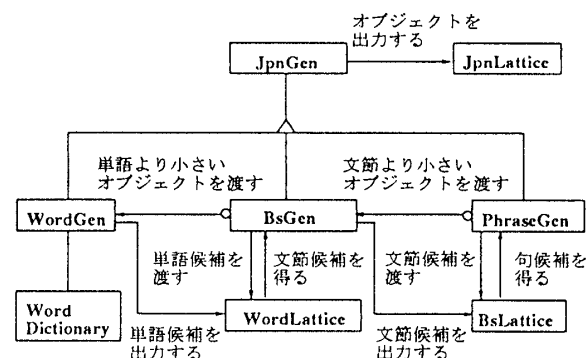


図3: ジェネレータクラス

プリメントはC++言語で行ない、現在UNIXワークステーションおよびパソコンで稼働している。

4 おわりに

日本語解析系をオブジェクト指向に基づいてモデル化し、C++のクラスライブラリとして実現することによって、汎用性、再利用性の高いソフトウェアライブラリを構築することができた。

今後は、日本語解析関連の各種アプリケーション開発への適用を推進し、その結果を基に、ライブラリの充実を図っていく予定である。

参考文献

- [1] 小松順子. コスト最小法に基づく逐次確定型・形態素解析. 情報処理学会第47回全国大会, 1993.
- [2] J. ランボー他. オブジェクト指向方法論 OMT. トッパン, 1992.