

2007年10月30日

日本語形態素解析システム 「茶釜」の開発を振り返って

松本裕治

奈良先端科学技術大学院大学(NAIST)

情報科学研究科

「茶釜」とは

◆ 日本語形態素解析システム

- 日本語文を単語に分ち書きし，各語の品詞，活用形，原型，読みなどを出力
- 日本語解析の基本ソフト
- 隠れマルコフモデルに基づき，解析済みデータより確率パラメータを学習
- 品詞の連接確率，語の出現確率（それぞれコストとして実現）を用いた有限オートマトンとして実現し，高速化（新聞の年間記事を通常のパソコンで約3分で解析）

茶釜の実行例

The screenshot shows the ChaSen software window with the following content:

私はその人を常に先生と呼んでいた。

私	ワタシ	私	名詞-代名詞-一般		
は	ハ	は	助詞-係助詞		
その	ソノ	その	連体詞		
人	ヒト	人	名詞-一般		
を	ヲ	を	助詞-格助詞-一般		
常に	ツネニ	常に	副詞-一般		
先生	センセイ	先生	名詞-一般		
と	ト	と	助詞-格助詞-一般		
呼ん	ヨン	呼ぶ	動詞-自立	五段・バ行	連用タ接続
で	デ	で	助詞-接続助詞		
いた	イタ	いる	動詞-非自立	一段	連用形
た	タ	た	助動詞 特殊・タ		基本形
。	。	。	記号-句点		
EOS					

VisualMorphs: 茶釜の内部状態の可視化ツール

VisualMorphs -p property.vm -a analyzer.vm -c C:\WINDOWS\デスクトップ\test.txt

ファイル PartOfSpeech Inflection

全文解析 ▲ ▲ 5 じゃあ京都行くまでに通行止めとかないのかなあ

部分解析 × ● 単語に区切られていない

切り出し ▼ ▼

見出し語 品詞 活用 最小コスト

基本形 全文コスト 10015.0

読み 全文解析幅 5000.0

発音 部分解析幅 10000.0

破棄

BOS/EOS 0
 単語 名詞 一般 3929
 に 助詞 格助詞 一般 0
 区切ら 動詞 自立 2950
 れ 動詞 接尾 0
 て 助詞 接続助詞 6
 い 動詞 非自立 0
 ない 助動詞 0
 BOS/EOS 0

単 接頭詞 名詞 接続 2315
 に 助詞 副詞化 0
 区 名詞 一般 3556
 切ら 動詞 自立 1980
 て 動詞 非自立 1445
 い 動詞 自立 1344

単 名詞 一般 3755
 語 名詞 接尾 一般 2044
 区 名詞 接尾 助数詞 2097
 語 名詞 2045

簡単な自己紹介(と主な言語処理ツール)

- ◆ 1979.4: 電子技術総合研究所入所
 - 構文解析システムBUPの開発
- ◆ 1984.9-1985.7: 英国Imperial College
 - 並列構文解析システムSAXの設計
- ◆ 1985.9-1987.11: 新世代コンピュータ技術開発機構(ICOT)
 - 並列論理型言語による並列プログラミング
 - 論理型言語による副作用・後戻りなしの動的計画法の実装
 - ◆ 並列構文解析(SAX, PAX), 形態素解析(LAX)の実装の基盤
- ◆ 1988.10: 京都大学長尾研究室
 - 形態素解析システムJumanの開発
- ◆ 1993.4: 奈良先端科学技術大学院大学
 - 茶筌, その他の言語処理ツール

言語処理ツールの開発と公開について

- ◆ 構文解析システムBUP
- ◆ 並行構文解析システムSAX
 - 電総研時代の田中穂積室長, ICOT時代の淵一博所長の勧めにより, フリーソフトとしてコンパイラを公開
 - ICOTでは, 開発した種々のツールやデータをICOT free softwareとして無償公開
- ◆ 形態素解析システムJuman
 - 京大長尾真先生の勧めにより開発・公開
 - ◆ 辞書システム: 妙木裕氏(当時B4)
 - ◆ C言語による実装: 黒橋禎夫氏(当時D1)
- ◆ 奈良先端大でもすべての言語処理ツールをフリーソフトとして公開することを前提で開発

「茶釜」の開発経緯

- ◆ Jumanのパラメータの機械学習化および辞書システムの改良等による高速化
 - 1996年夏のプロジェクトとして実施
 - ◆ 北内啓, 山下達雄, 今一修, 今村友明
 - 約10倍以上の高速化
 - 生駒市高山町の名産に合わせて「茶釜」と命名
- ◆ RWCPコーパスによる機械学習
 - そのためのコーパスの誤り修正
 - ◆ 1997年夏: 数万箇所 of 誤り修正
 - 新しい品詞体系へ変更: IPADIC作成
 - ◆ 1998年春: 学習に基づく最初の辞書IPADIC 0.8公開
 - ◆ 1999年暮: 浅原氏による学習プログラムに基づくIPADIC 2.0公開

茶釜の追加機能あれこれ

- ◆ 入出力機能の充実
 - 北内啓氏
- ◆ 辞書項目への発音情報の追加
 - 山田篤氏, 伊藤克亘氏
- ◆ 解析過程の可視化ツール
 - 美茶: 山下達雄氏
 - VisualMorphs: 松田寛氏
- ◆ Windowsへの移植(現在は, UNIX版と共通)
 - 平野善隆氏, 松田寛氏
- ◆ 部分解析機能
 - 高岡一馬氏
- ◆ 解析結果の管理(検索, 誤り修正)システム: 茶器
 - 松本裕治他

本当は何が大変だったか(その1)

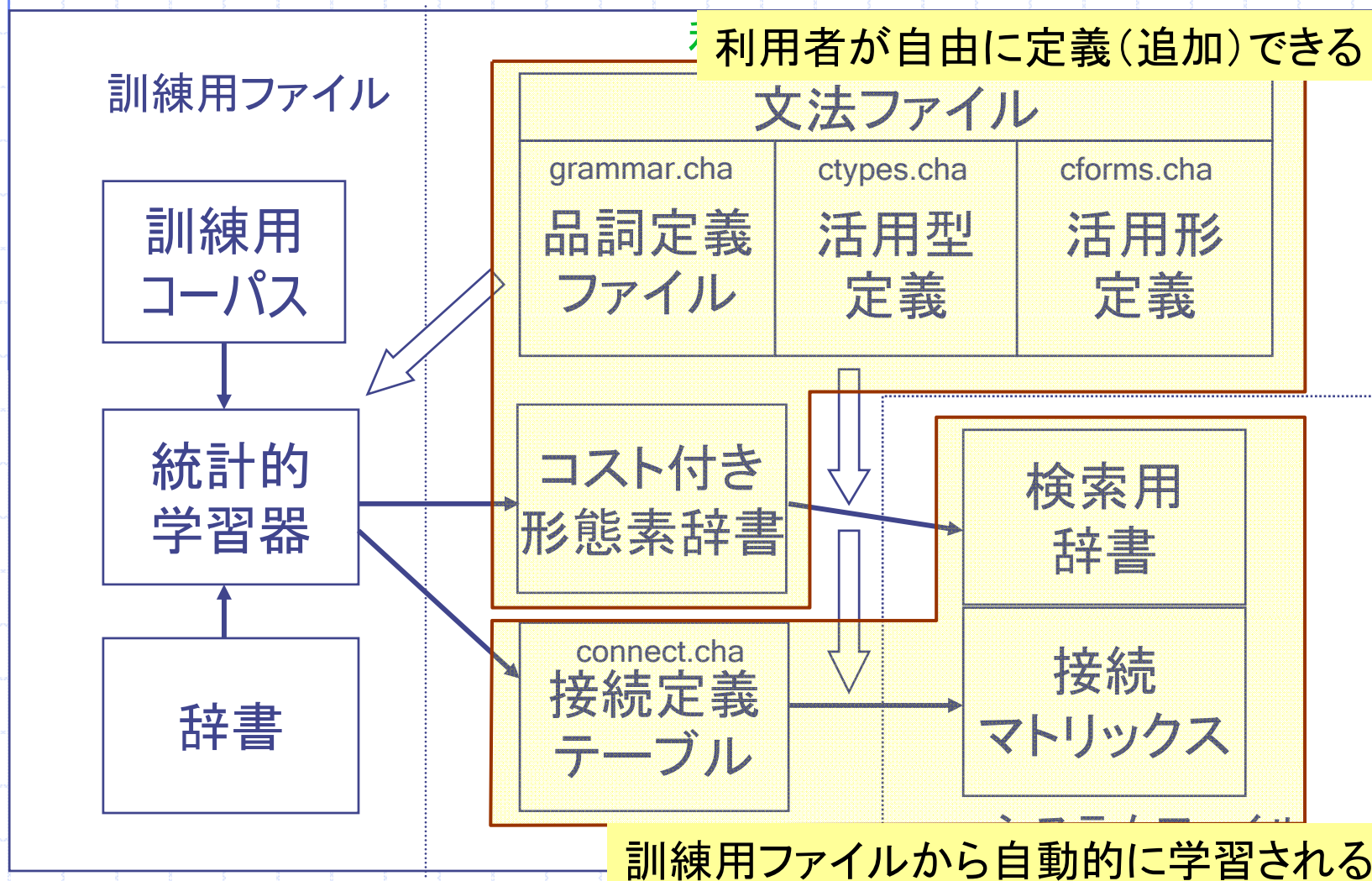
◆ 辞書 — 自由に使える辞書の不足

- 当初: Wnnの辞書を利用
- その後: ICOT fee softの岩波国語辞典を追加
- 人名, 地名, 組織名の追加

◆ 辞書システム

- B-tree (妙木): 辞書本体がファイルシステムにあると仮定
- NDB (黒橋): UnixのDBシステムによる擬似的なTRIE実装
- Patricia木 (米沢恵司, 山下): 辞書をメモリ上に展開
- 二重配列TRIE (工藤拓)

茶釜の辞書システムの構成



本当は何が大変だったか(その2)

- ◆ 機械学習を行うためのタグ付きコーパス
 - 1996年暮に公開されたRWCPコーパス
 - ◆ 多数の不整合が内在
 - 1997年夏: 学習前後の結果を比較し, 変換(修正)パターンを抽出するツールを作成し, 不整合や誤りの修正
 - タグ付きコーパス内および辞書との不整合を図ることの困難
- ◆ その後の辞書システムやコーパス管理ツールの開発につながる
- ◆ 実は, 茶釜本体の開発よりも周辺の整備に遥かに多くの労力と時間を費やした

我々のグループで構築し公開している 言語処理ツール

◆ 言語解析ツール

■ 日本語形態素解析システム

- ◆ 茶釜(ChaSen) [Asahara & Matsumoto 00] 日本語版, 中国語版, (英語版)
- ◆ MeCab [Kudo 04]

■ 日本語係り受け解析: 南瓜(CaboCha) [Kudo 02]

- ◆ 英語, 中国語単語係り受け解析 [Yamada 03, Chen 04]

■ 汎用チャンカー: YamCha [Kudo 01]

■ 未知語識別器: bar [Asahara 04]

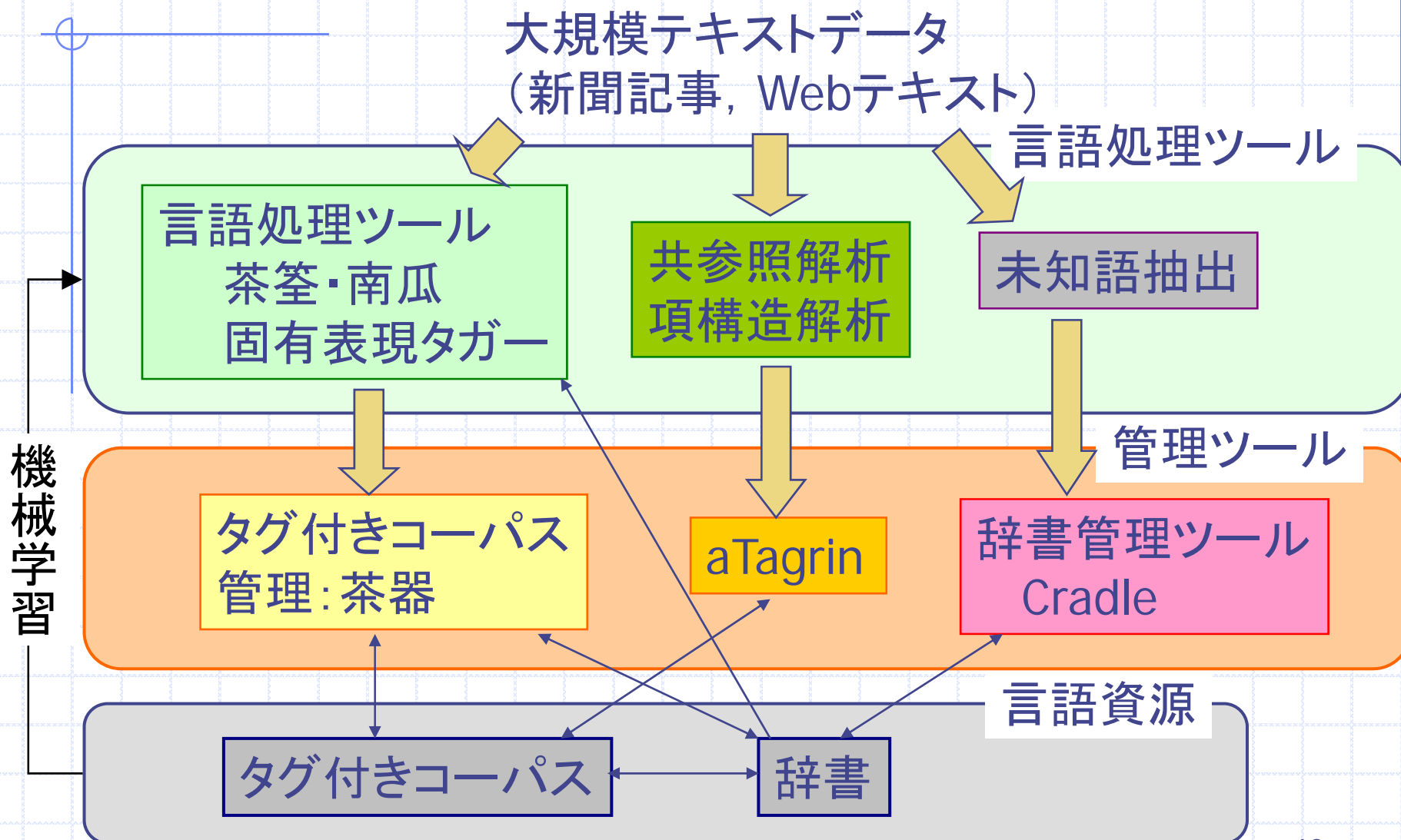
◆ 言語データ管理ツール

■ タグ付きコーパス管理ツール: 茶器 [Matsumoto 06]

■ 辞書管理ツール: Cradle

■ 汎用アノテーションツール: aTagrin

言語処理ツールと コーパス管理システムの関係



様々な人々の支え

- ◆ 言語処理および成果としてのソフト公開の重要性
 - 淵一博(元ICOT所長), 田中穂積(現中京大学), 長尾真(現国会図書館館長)
- ◆ 言語処理ツール研究・開発に関する協力
 - 伝康晴(千葉大学), 黒橋禎夫(京大), 妙木裕(キヤノン), 山田篤(ASTEM), 宇津呂武仁(筑波大学), 浅原正幸(NAIST), 今一修(日立), 竹内孔一(岡山大学), 山下達雄(Yahoo), 平野善隆, 北内啓(NTTデータ), 米沢恵司(野村総研), 松田寛, 高岡一馬(ジャストシステム), 工藤拓(Google), その他研究室の多くの学生諸君
- ◆ 様々なユーザからのコメント, フィードバック