日本語形態素解析システム 茶筌

山下 達雄 (tatuo-y@is.aist-nara.ac.jp)
奈良先端科学技術大学院大学 情報科学研究科

茶筌は、奈良先端科学技術大学院大学松本研究室で開発されたコスト最小法による日本語形態素解析システムである。JUMAN2.0[2] が原形となっている。

広く一般に利用されることを目的としてフリーの形態素解析システムとして公開されている。機械翻訳や対話処理などの自然言語処理システムの前処理に利用されているだけでなく、freeWAIS などの日本語全文検索ソフトのインデキシングなどにも利用されている。UNIX 版だけでなく、Windows 95 版も提供されており、国文・教育関連の文系の研究者にも利用されている。また、付属の約 20 万エントリの形態素辞書はフリーのものとしては最大規模で、他の形態素解析システムや、形態素解析以外の研究にも多く利用されている。

茶筌の特徴として、品詞体系、形態素辞書の項目、連接規則がユーザにより自由にカスタマイズできるという点があげられる。現在、公式にサポートされている品詞体系には益岡田窪文法と RWCP で採用されているもの (学校文法に準拠) がある。後者は現在辞書の整備を行なっている最中で、近日中に公開予定である。

また、文法だけでなく、コスト最小法で利用する、形態素コスト・連接コストを辞書や連接規則定義ファイルで設定することができる。これにより、コーパスから統計的に学習した確率値をコストに変換して用いれば、茶筌を確率モデルに基づく形態素解析システムとして利用できる。

グラフィカルユーザインターフェース (図 1) の開発や可変長連接規則の実装などの機能拡張が行われており、今後も継続されていく予定である。

茶筌に関する詳しい情報はマニュアル [1] を最新の動向に関しては茶筌ホームページ[1]を参照されたい。

参考文献

[1] 松本裕治, 北内啓, 山下達雄, 平野善隆, 今一修, 今村友明. “日本語形態素解析システム『茶筌』version1.0 使用説明書”, NAIST Technical Report, NAIST-IS-TR97007, February 1997.

[2] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真. “日本語形態素解析システム JUMAN 使用説明書 version 2.0”, NAIST Technical Report, NAIST-IS-TR94025, July 1994.

[1] http://cl.aist-nara.ac.jp/lab/nlt/chasen.html

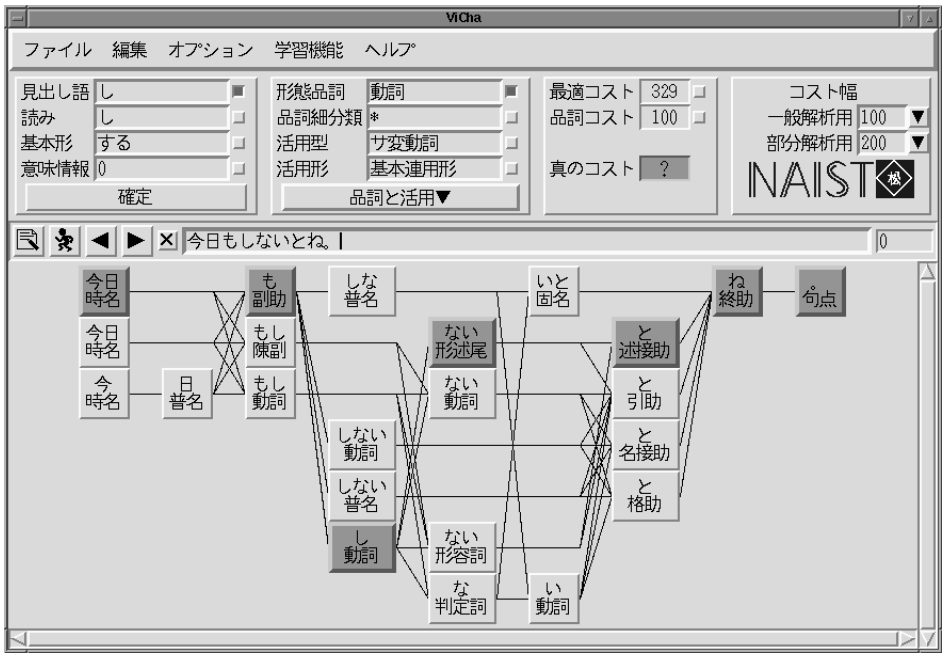

図 1: 茶筌の解析結果を表示する GUI