統合データベースプロジェクトシンポジウム
(June 12, 2009)

専門用語辞書管理システム および 用語の内部構造アノテーションについて

山田恵美子, 呂嘉, 松本裕治
(東京大学) (奈良先端科学技術大学院大学)

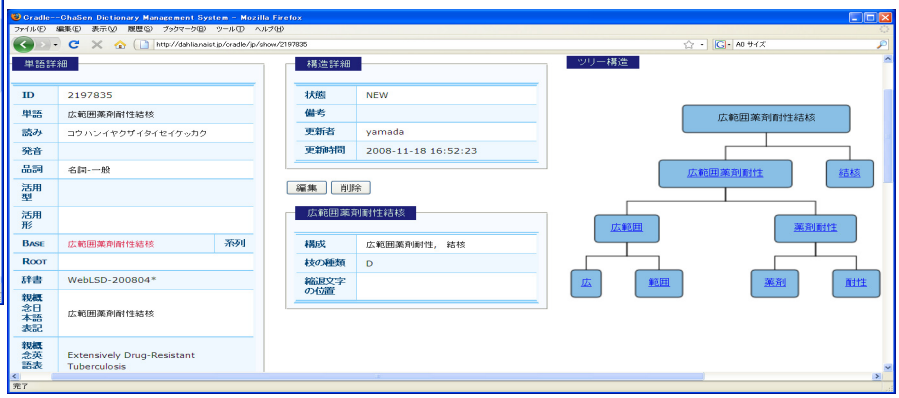
概要:医学生物学関連の専門用語の表記、読み, 品詞, 英訳, シソーラスコード等の意味クラス情報などの基本情報, および, 同義語へのリンクや内部構造を記述できる辞書管理システムを開発している. 現在は, ライフサイエンス辞書の日本語の用語が対象. 複合語として複雑な構造をもつ専門用語の内部構造を単語単位の係り受け木構造としてアノテーションする機能を備えている. 一部の用語の内部構造を本システムを用いてよってアノテーションし, それを学習データとして用いて, 未解析の専門用語の内部構造の自動アノテーション実験を行った結果を報告する.

辞書管理システム「Cradle」

- ライフサイエンス辞書(京大金子研 2008.4版)を形態素解析用辞書として管理
- 見出し, 読み, 品詞などの用語の基本情報, および MeSHのシソーラスコード, および, 英訳の情報を格納
- 専門用語(複合語)の内部構造アノテーション機能と内部構造表示機能
- 検索機能(綴り, 読み, 品詞, 構成語による検索)
- 同義語情報の付与

複合語の内部構造アノテーション画面

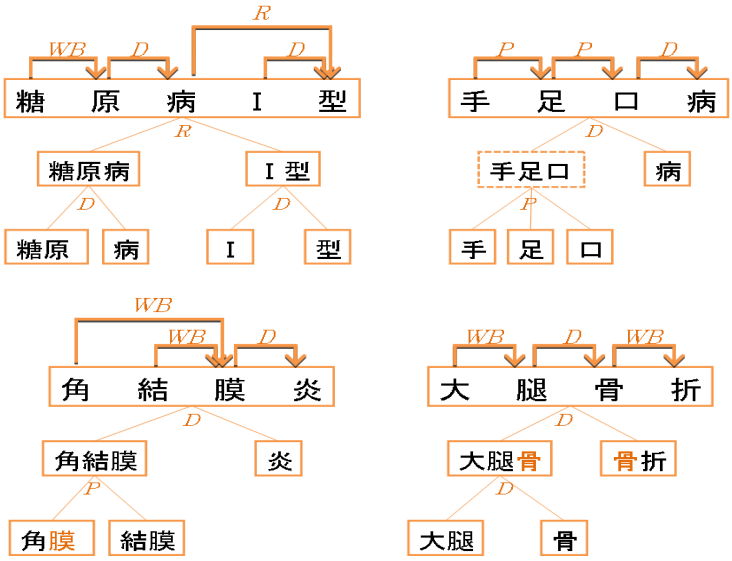
「結核」を含む語の検索結果画面



用語の詳細情報の提示画面(右は内部構造表示)

専門用語の内部構造解析

- 複合語の構造は単語からなる係り受け木によって記述可
  - D(通常の係り受け), R(逆向き係り受け), P(並列), U(その他の関係)によって表現
- 共通の文字が省略される(縮退する)現象がある
  - これを表現するため, 文字間の係り受けによって解析
  - WB(語の構成要素の先頭として係る), WI(語の構成要素の内部として係る)の関係を解析用に新たに定義



文字単位の係り受け解析(用語上部の矢印)と
構成要素となる単語による木構造表現(下部の木構造)

専門用語の内部構造解析実験

- 一部の用語に単語間関係の内部構造アノテーションを行い, これを学習データとして自動解析実験を行った.
  - データ
    - ライフサイエンス辞書の掲載語から804語
  - 手法
    - Shift-Reduce法(MaltParser)[Nivre 2004]を利用
  - 結果:(5分割交差検定による精度)
    - 文字間係り受け 96.9% (ラベル無し:97.8%)
    - 用語全体の解析 87.6% (ラベル無し:89.9%)

[Nivre 2004] Nivre, J. and Scholz, M. "Deterministic Dependency Parsing of English Text," In Proceedings of COLING 2004, pp.64-70, Geneva, August 2004.