資料２－６

2011年2月21日（月）
統合DBプロジェクト合同会議資料

「専門用語辞書管理システムと専門用語解析技術の開発」

奈良先端科学技術大学院大学
松本裕治

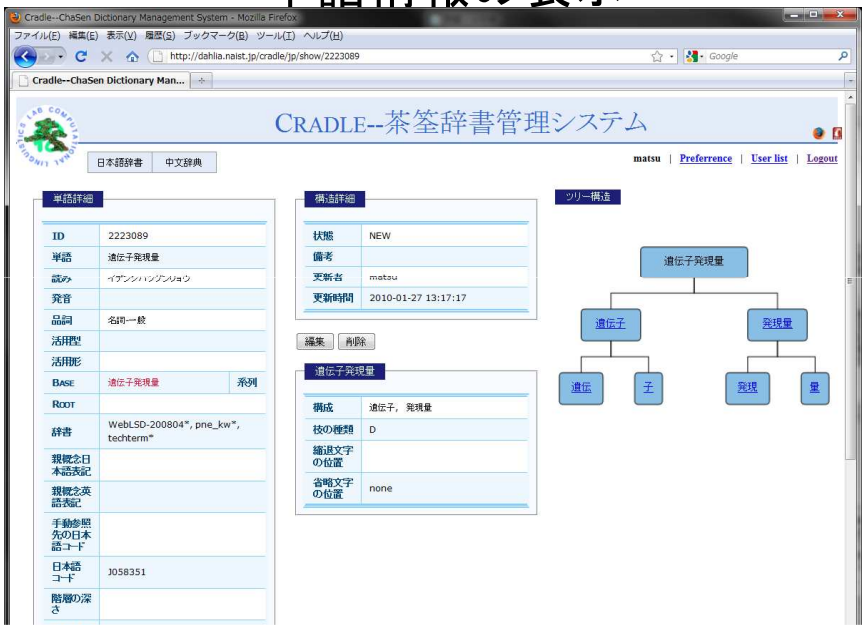
1. 専門用語辞書システムの設計
2. 専門用語解析技術の開発
3. 専門用語抽出ツールの設計と開発

研究項目と成果

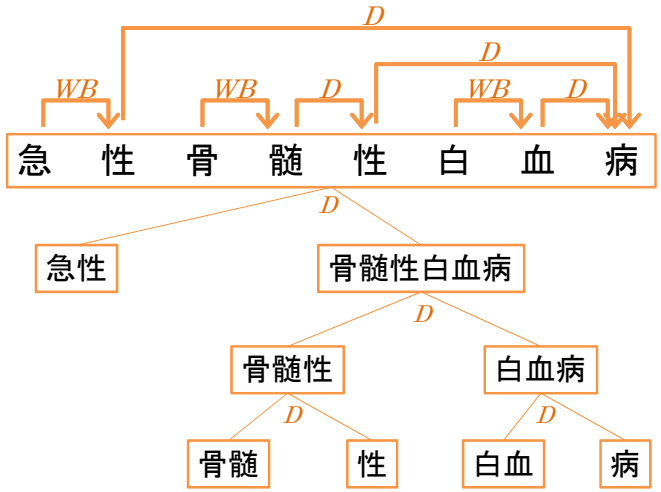
1. 専門用語辞書システムの設計
   - 95000語のライフサイエンス辞書を対象に，新規用語の登録，同義語情報の記述，種々の検索や用語の語構成（内部構造）情報を格納することができる専門用語辞書システムの開発を行った．
2. 専門用語解析技術の開発
   - 専門用語辞書システムの機能を利用して，約3500語の用語の内部構造付与作業を行った．これを利用して，専門用語の内部構造の自動解析技術を開発した．
3. 専門用語抽出ツールの設計と開発
   - 文書中の用語の検索と，その用語のシソーラス内での位置を表示するシステムを開発した．また，シソーラスにはない新規の用語に対して，類似の用語とそのシソーラス上での位置を示すことにより，新規用語のシソーラス登録を支援するシステムを開発した．

1．専門用語辞書システム：検索画面

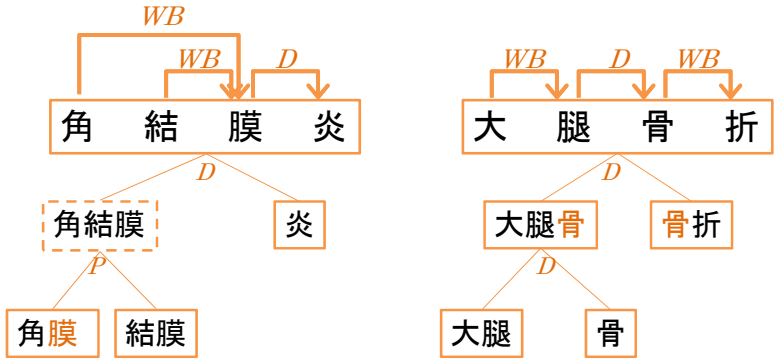
単語情報の表示



2. 専門用語解析技術
文字単位の係り受けによる用語の内部構造解析



2. 専門用語解析技術
文字単位の係り受けによる用語の内部構造解析



3. 専門用語抽出ツール

トップページ（http://cl.naist.jp/kazuo-h/）:

トップページ　バイオ医療専門用語の検索　類似文脈の検索

バイオ医療専門用語と類似文脈の検索

コーパスとして PNE (蛋白質・核酸・酵素)を，シソーラスとして LSD (ライフサイエンス辞書)（2008年度版）を用い，次の2つの検索を行います．

- PNEとLSD に対するバイオ医療専門用語の検索（クエリ＝専門用語）
- PNE に対する類似文脈の検索（クエリ＝文脈＋専門用語）

PNE に対する専門用語の検索には，sary を用います．類似文脈の検索では，クエリとして与えられる文脈，および，PNEに登場する専門用語の周辺文脈を，MeCab と CaboCha を用いて抽出し，それら文脈間のコサイン類似度を計算します．

将来的にはシソーラスに未登録の専門用語（新しい治療方法等）の新規登録を支援することが目標であり，より精度の高い類似度計算方法を現在開発中です．

(i) バイオ医療専門用語の検索

検索画面:

トップページ　バイオ医療専門用語の検索　類似文脈の検索

バイオ医療専門用語の検索（クエリ＝専門用語）

クエリ: 急性骨髄性白血病 Submit query

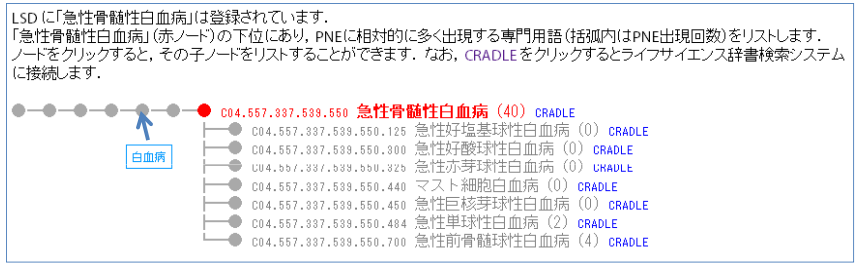
検索結果（PNE検索）:

「急性骨髄性白血病」を含む文脈は，PNEに10件以上あります．

前者は成熟傾向を示す 急性骨髄性白血病 (AML) を表現型とする
最近では，X染色体DNAの多型性を指標として 急性骨髄性白血病，子宮平滑筋腫，腎に発生するウィルムス腫瘍が単クローン性で
たとえば，急性骨髄性白血病 細胞では34人の患者中19例にp53蛋白質が増加することが見い
急性骨髄性白血病 患者 (24歳，男性) 末梢血
急性骨髄性白血病 株細胞としてML-1以外にKG-1も広く使われている
急性骨髄性白血病 の腫瘍性芽球も単クローン性である
先述のとおり，急性骨髄性白血病 の一部の症例では，赤血球や顆粒球も腫瘍性芽球細胞と同じクロー
これから述べるHL-60細胞は，急性骨髄性白血病 患者 (FAB分類M_2) (以前はM_3とされていたが，近
この後遺症の期間は，同年齢で原爆被曝した人の 急性骨髄性白血病 の発生期間 (図7) とほぼ一致する
図7は，Ichimaruらが 急性骨髄性白血病 (AML) と急性リンパ性白血病 (ALL) に分けて発表し

(i) バイオ医療専門用語の検索

検索結果(ライフサイエンス辞書のシソーラス表示):

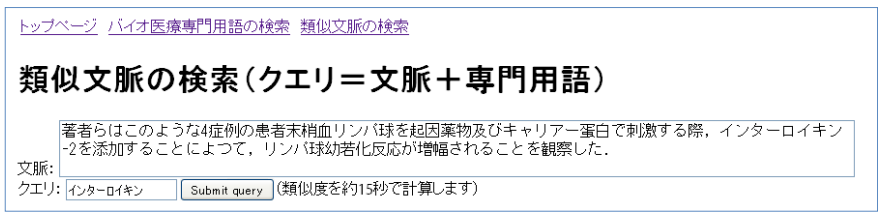


- Cradle-LSD(ライフサイエンス辞書検索システム)とリンク: 単語の詳細情報が表示される
- ノードをクリックすると, その子ノードの一覧を表示する(次のスライド参照).

(ii) 類似文脈の検索

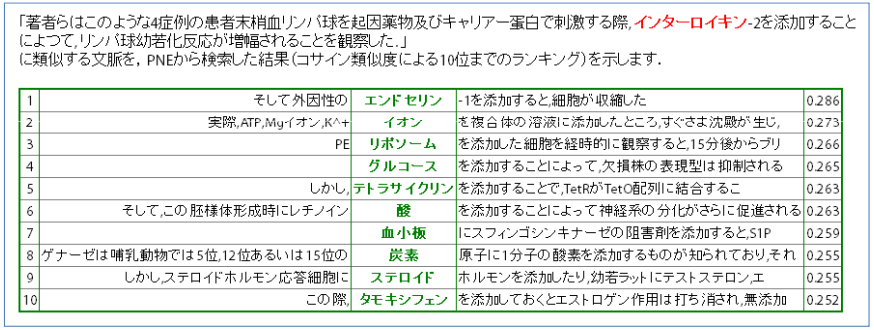
目的: クエリとして与えられる文脈, および, PNEに登場する専門用語の周辺文脈を抽出し, それら文脈間の類似度を計算する.

検索画面:



(ii) 類似文脈の検索

検索結果(対象の用語と類似の用語を表示)



1	そして外因性の	エンドセリン	-1を添加すると,細胞が収縮した	0.286
2	実際,ATP,Mg-イオン,K^+	イオン	を複合体の溶液に添加したところ,すぐさま沈殿が生じ,	0.273
3	PE	リボソーム	を添加した細胞を経時的に観察すると,15分後からプリ	0.266
4		グルコース	を添加することによって,欠損株の表現型は抑制される	0.265
5	しかし,	テトラサイクリン	を添加することで,TetRがTetO配列に結合するこ	0.263
6	そして,この胚様体形成時にレチノイン	酸	を添加することによって神経系の分化がさらに促進される	0.263
7		血小板	にスフィンゴシンキナーゼの阻害剤を添加すると,S1P	0.259
8	ゲナーゼは哺乳動物では5位,12位あるいは15位の	炭素	原子に1分子の酸素を添加するものが知られており,それ	0.255
9	しかし,ステロイドホルモン応答細胞に	ステロイド	ホルモンを添加したり,幼若ラットにテストステロン,エ	0.255
10	この際,	タモキシフェン	を添加しておくとエストロゲン作用は打ち消され,無添加	0.252

クエリおよび，クエリと類似する専門用語のLSDにおける位置

プロジェクト終了後の維持・活用について

- 専門用語辞書システム
  - 以下のサイトで継続して運用
    - http://dahlia.naist.jp/lsd
  - 検索は誰でも実行可能
  - 登録者は，新規用語の追加，および，用語の内部構造情報やその他の情報の付与を行うことが可能
- 専門用語抽出ツール
  - 以下のサイトで継続して運用
    - http://cl.naist.jp/kazuo-h/
  - 「蛋白・核酸・酵素」の検索や類似語検索は継続して運用
- これらのツールの他サイトへの移植についても検討