統合データベース開発

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　新保 仁　浅原 正幸

資料3－5

平成 19 年度業務報告

❶ 専門用語辞書システムの設計

❷ 専門用語解析技術の開発

❸ 専門用語タグ付け手法の設計

キーワード: 専門用語

多くは複合語

構成要素間の複雑な関係

➔ 内部構造を考慮に入れた辞書管理システム

➔ 構成要素間の関係解析

並列句として用いられた場合に部分的な省略を受けやすい

➔ 並列句解析精度の向上

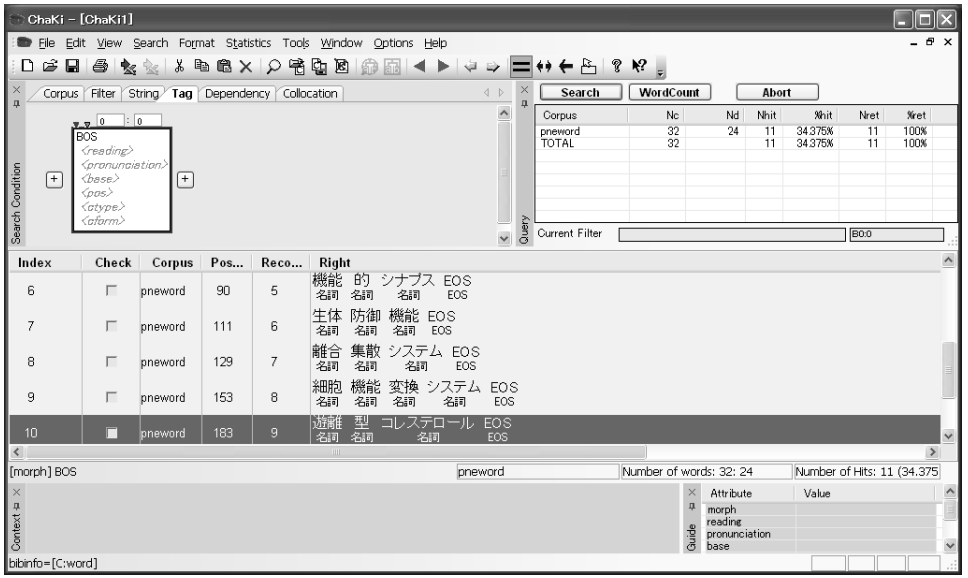
❶ 専門用語辞書システムの設計

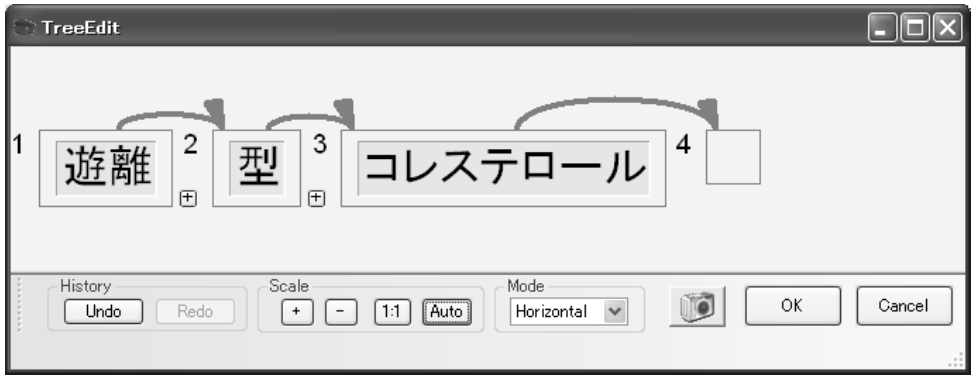
専門用語の意味的類似性，同義語を判定する基礎技術を開発し，用語の意味関係を記述することのできる用語辞書システムの基本設計を行う．

用語辞書管理システム (プロトタイプ)

用語構成要素間の係り受け関係のタグ付けGUI を備える

係り受け強さの自動学習（次年度以降）を行う際に必要となる，訓練データ作成に使用可能





❷ 専門用語解析技術の開発

専門分野のテキストには複雑な専門用語が数多く出現する．特に，複数の類似の用語が並列表現として出現する場合や部分的に省略が行われている場合など，用語辞書の充実だけでは完全な用語抽出を行うことができない．これに対処する解析技術として，専門用語の並列構造解析技術の開発を目指す．

専門用語と並列句

しばしば部分的な省略が起こるため，正確な解析のためには辞書の整備だけでは不十分

erythroid, myeloid and lymphoid cell types

= erythroid cell type and myeloid cell type and lymphoid cell type

human T and B cells

= human T cells and human B cells

recombinant human nm23-H1, -H2, mouse nm23-M1, and -M2 proteins

= ???

並列句解析

並列句の構成要素には構文的な類似性が高いことが多い

➔ 文中の類似部分を検出

既存法

類似性尺度＝ヒューリスティックなスコアリング規則 (固定)

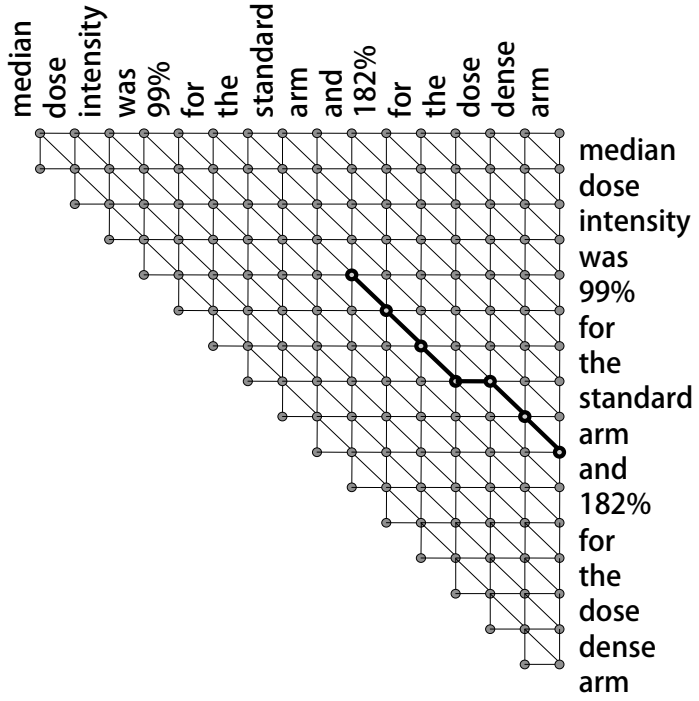
一般の日本語文（新聞記事等）の解析のためにチューニング

問題点

医学生物学分野 ≠ 新聞記事・英語 ≠ 日本語

開発した手法

機械学習技術を用いて分野・言語に則したスコアづけを学習



❸ 専門用語タグ付け手法の設計

専門用語解析の基礎データの蓄積を行うため，専門分野テキストに対して，用語の出現箇所の特定や用語の種類の分類を行うため，タグ付けの表現法，および，タグ付け手法を設計する.

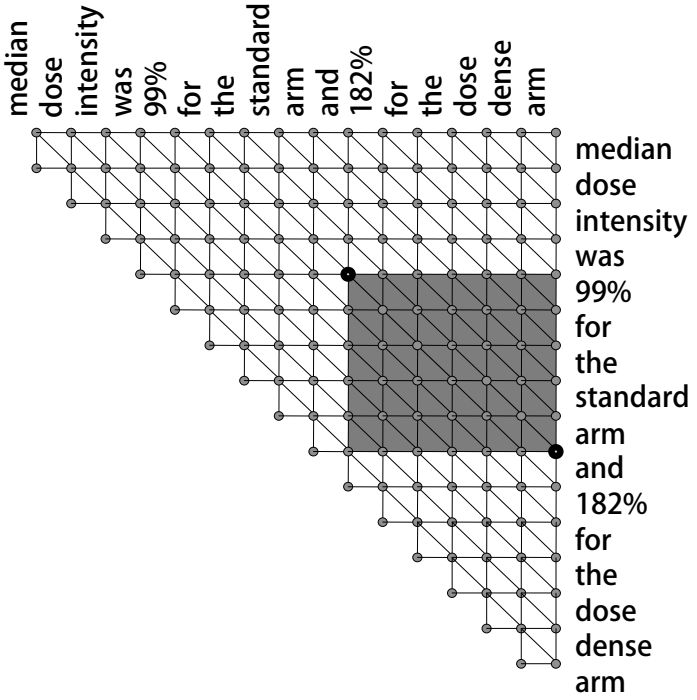
並列句タグ付け

教師つき機械学習＝タグ付け正解データが必要

単語単位のタグ付け=コスト (時間・手間) 大

並列句範囲のタグ付け＝コスト比較的小

➔ 範囲のみを指定して，学習可能な並列句解析法を開発



平成 20 年度計画概要

専門用語辞書システム

今年度開発した辞書システムプロトタイプを用いて，辞書管理の容易さの検証

大久保・川本先生よりシード辞書受け取り単語登録

並列句解析処理

日本語医学生物学テキストへの適用

精度向上のための有効な素性の発見

リソース（訓練データ，タグ付け）の蓄積

専門用語抽出および内部構造の解析

今年度開発した GUI を用いて単語登録，用語内部構造のタグ付けを行う

➔ 専門用語自動抽出および係り受け解析器用の訓練データとして蓄積

機械学習によりコーパスから半自動で用語収集・用語内係り受け構造解析を行う