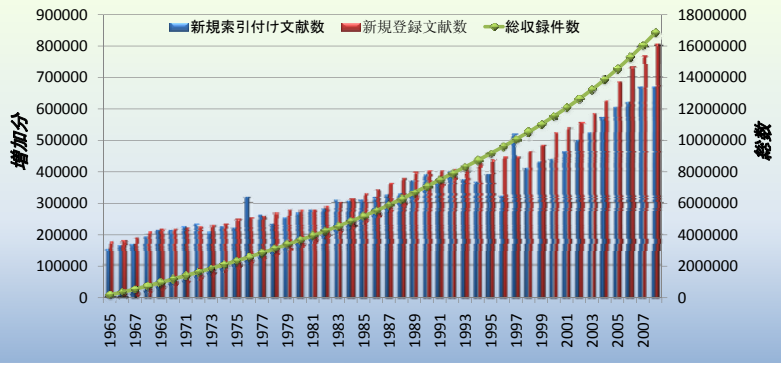
生命科学文献からの情報抽出とテキストマイニング技術の開発

東京大学大学院情報理工学系研究科
大田朋子・宮尾祐介・松崎拓也・Brian Kemper・辻井潤一

研究の背景と目的

- 生命科学の知識は論文（テキスト）として蓄積されている
- 生命科学論文の発表数が爆発的に増大している
- 膨大な知識が未整理のまま論文に埋もれている

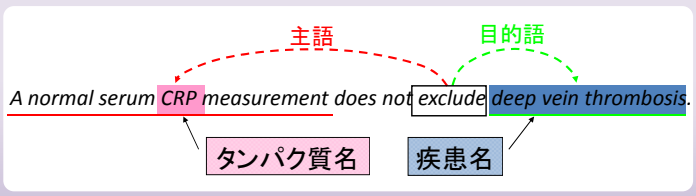
⇒ 膨大な論文に埋もれた知識を効率よく活用したい
＝ テキストマイニング



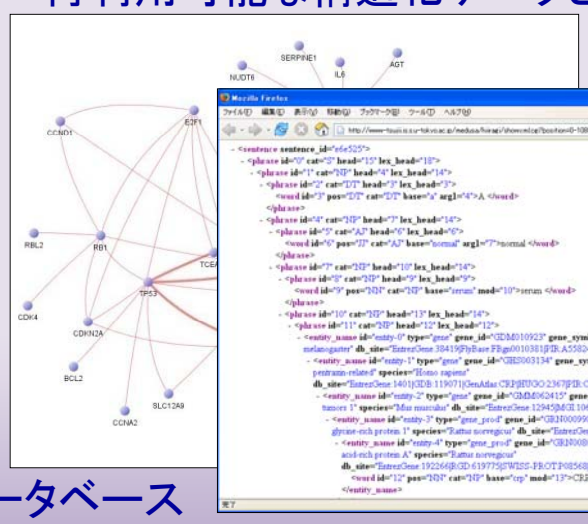
自然言語処理技術の生命科学文献への適用

テキストデータ（PubMed）

構文解析，用語認識，PPI認識，etc.



再利用可能な構造化データとして蓄積



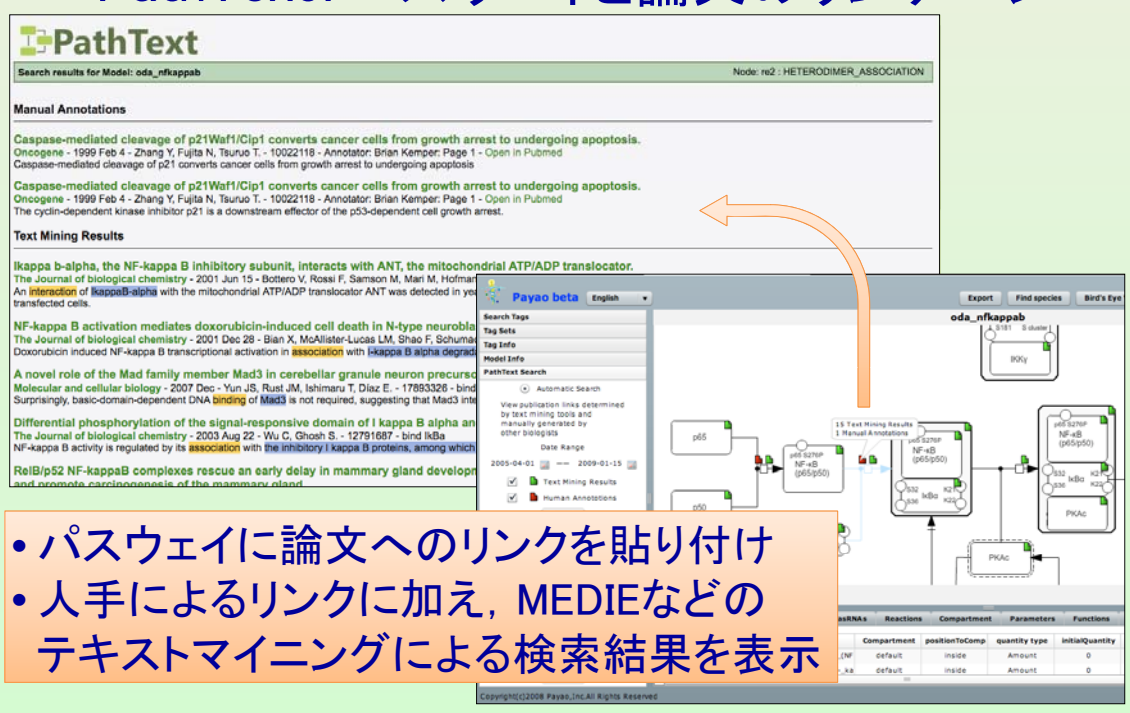
関係データベース

構造化テキストデータベース

MEDIE：意味に基づく文献検索エンジン

- 主語，述語，目的語を指定した検索
- オントロジーによる同義語（synonym）の検索
- 文の意図（目的，手法など）を指定した検索

PathText：パスウェイと論文のリンケージ

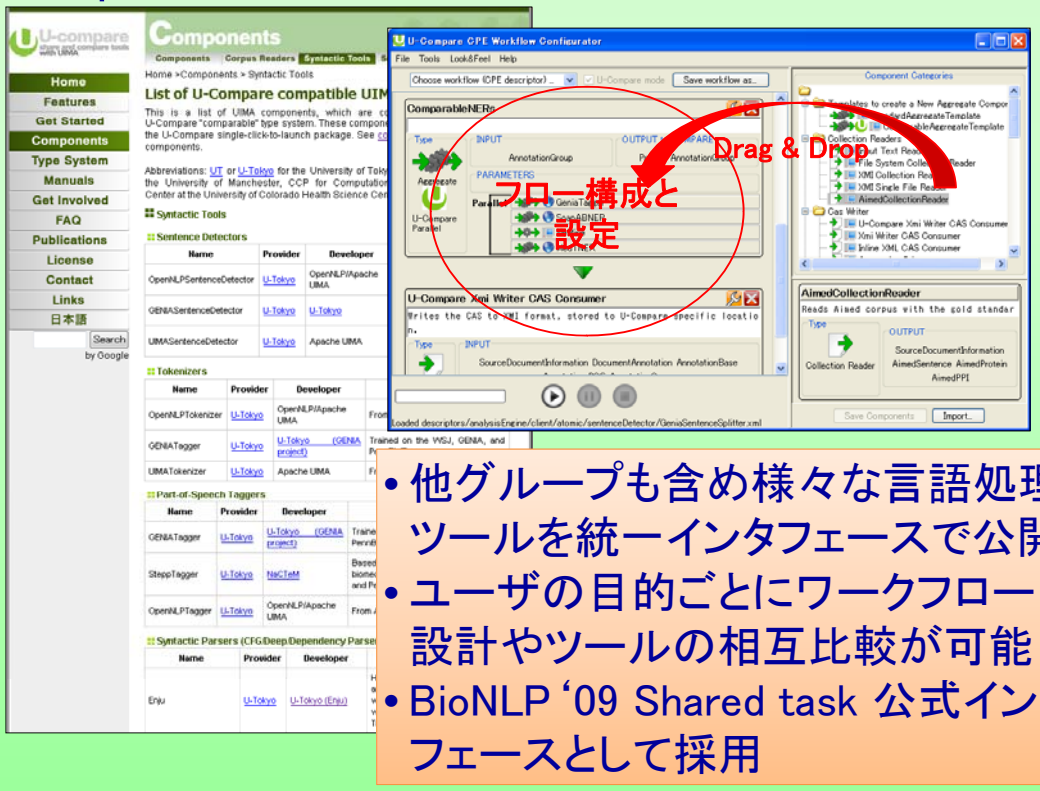


- パスウェイに論文へのリンクを貼り付け
- 人手によるリンクに加え，MEDIEなどのテキストマイニングによる検索結果を表示

Info-PubMed：二項関係のブラウザ

- 論文から自動獲得したタンパク質相互作用（PPI）や遺伝子疾患関係（GDA）を統一的にブラウズ

U-Compare：汎用テキストマイニングインタフェース



- 他グループも含め様々な言語処理ツールを統一インタフェースで公開
- ユーザの目的ごとにワークフロー設計やツールの相互比較が可能
- BioNLP '09 Shared task 公式インタフェースとして採用