# 専門用語の内部構造解析

著者:

山田恵美子(奈良先端科学技術大学院大学)

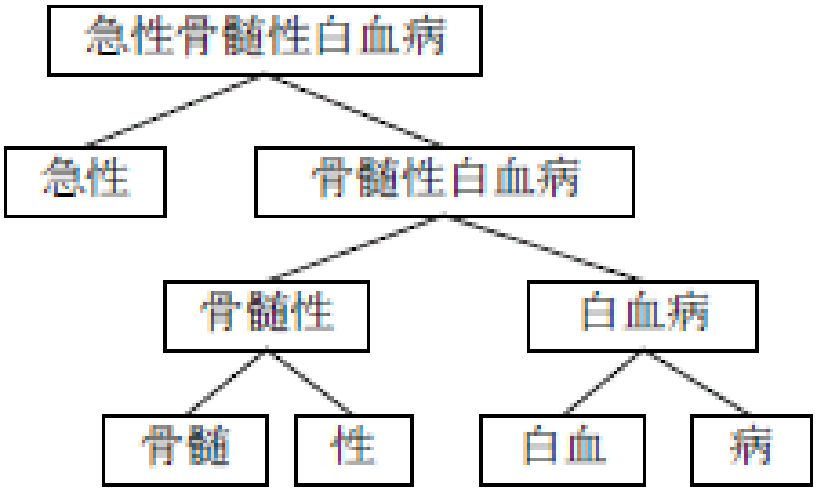
(東京大学大学院)

松本裕治　(奈良先端科学技術大学院大学)

発表者:三上健太

# はじめに

- 特定分野の文章において専門用語は特有の意味を保持
  → それを知ることは文章を解析する際に有用
- 専門用語は複合名詞が多い



# はじめに(2)

- しかし・・・

  専門用語は数多く存在、動的に作られうる

  →予め全ての語に内部構造を記述しておくのは現実的でない

- 本研究の目的

  ・専門用語の内部構造で見られる複雑な現象を調べる

  ・内部構造のタグ付けを行うための手法を提案する

- 一般の文で使われる係り受け構造だけでは表現できないものを表現するのに必要な表現方法を提案

- SVMによる自動解析の試みを報告

# 専門用語

- 対象は生命科学分野の専門用語(特に疾患名や解剖部位の用語の内部構造)
- 内部構造：その語を構成する構成要素の間の係り受け関係
- 専門用語

  - 特定の領域内で使われるもの
  - その分野の文化に依存して独自の内部構造を持ちうる

  ex.「糖原病Ⅰ型」:「Ⅰ型」が「糖原病」に係っている

  「角結膜炎」:「角膜」と「結膜」が並列に結びつき文字の縮退が起こっている

  → 従来の形態素への分割と係り受け構造をそのまま適用しても表現しきれない

# 専門用語(2)

- ある程度複雑な概念を限られた文字数で表す
  → 構成性が弱い → 内部構造を一意に決定するのが困難
  ex.「全前脳症」の「全」はどこにかかるか？
  「前脳」:ヒトの発生段階で脳の一部として存在し、複数の組織に分化する部位
  「全前脳症」:前脳が分化せずそのまま残ってしまい奇形が生じるという疾患
  →「前脳が全部そのまま残っている」ということ
  →「全」は「前脳」に係る
- 複合名詞の内部構造に関する研究がいくつかある
  → いずれも従来の形態素解析・係り受け解析の域を出てない
  → 本研究で扱う専門用語には不十分

# 内部構造の表現方法

- 複合語の内部構造：構成要素とその間の関係から成る
- 分割された構成要素間の関係を分類
  D：前から後ろへの係り受け(急性 => 肺炎：急性肺炎)
  R：後ろから前への係り受け(糖原病 <= Ⅰ型：糖原病Ⅰ型)
  P：並列(脊髄＋小脳：脊髄小脳変性症)
  U：上記以外の結びつき(B＋1＋6：B16メラノーマ細胞)
- 内部構造は枝に上記4種のラベルのいずれかを付与した係り受け木で表現できる
- 以下の3つの制約を利用
  ・自身よりも後の形態素に係る
  ・1つの形態素が複数の係り先を持たない
  ・交差が起きない

# 内部構造の表現方法(2)

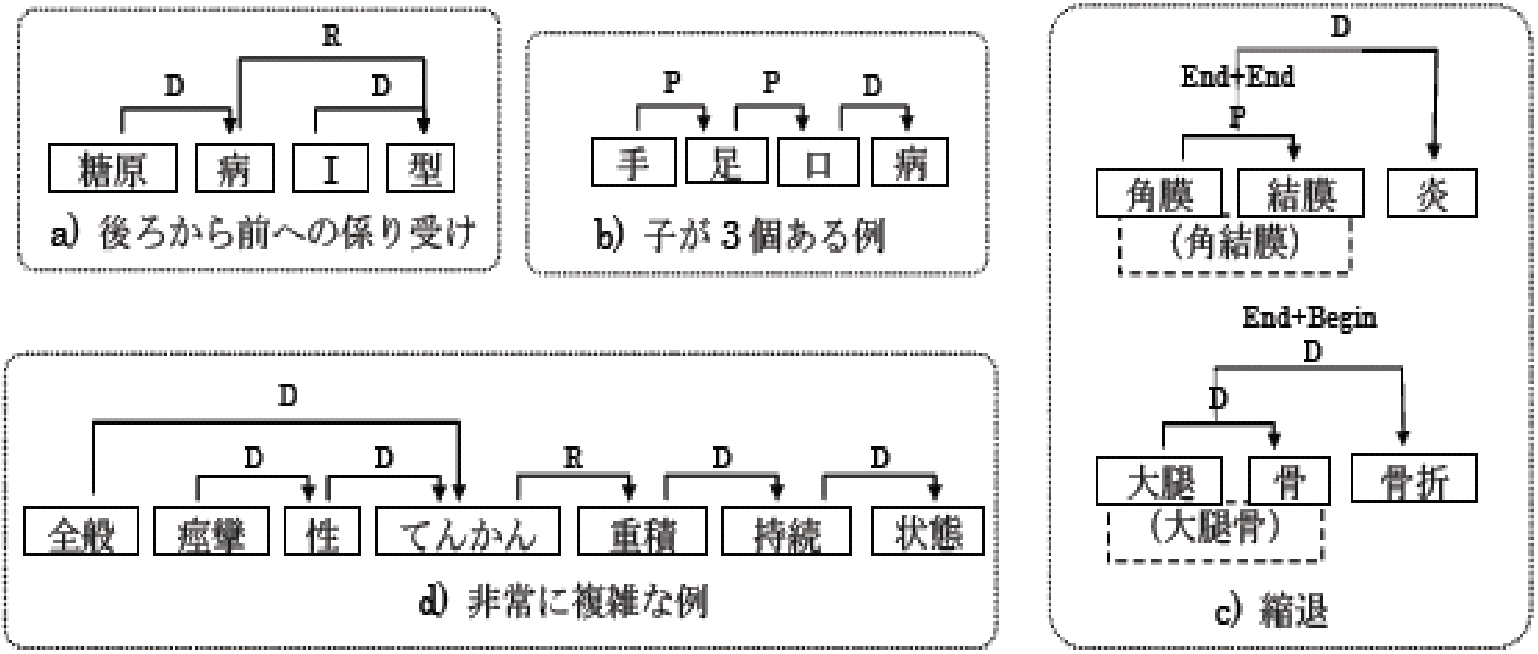
- 構文木は原則2分木
- ただし・・・

  ラベルがPまたはUの時は子ノードが2つ以上ある場合がある

  →同じ係り受け関係の連続で表現

- この表現方法は、今回対象としたデータに対して十分な表現力を持っている
- ラベルを導入したことによって、内部構造解析はラベル付きの係り受け解析として捉えられる

# ラベル付き係り受けによる内部構造表現



# 文字単位の係り受けによる表現

- 語を分割する際、文字の縮退を考慮する必要がある
- 縮退：複数の構成要素が結合する際、オーバーラップしている部分がまとめられる現象
- 縮退についてもラベルを付与する

  「End+End」：構成要素の最後の文字が縮退している

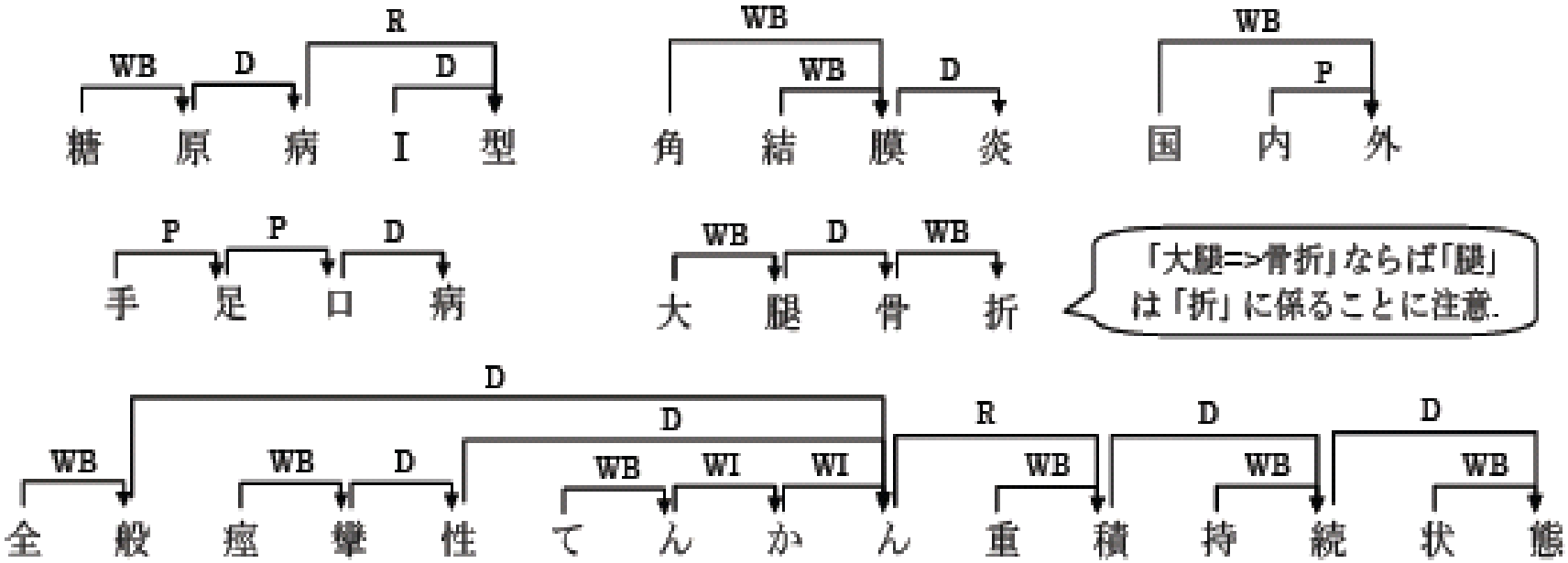
  「End+Begin」：構成要素の1番目の最後と2番目の最初の文字が縮退している

  「Begin+Begin」：最初の文字で縮退が起きている
- さらに構成要素そのものを作る係り受けを追加する

  WB：構成要素の先頭部分

  WI：それ以外の部分での係り受け

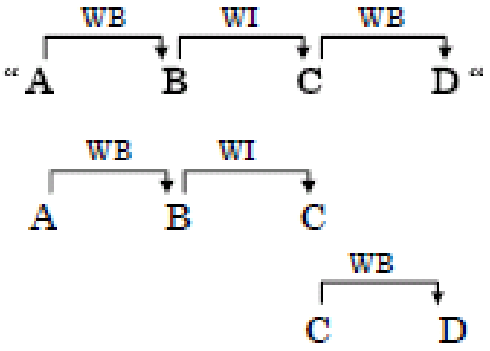
# 文字単位のラベル付き係り受けによる内部構造表現



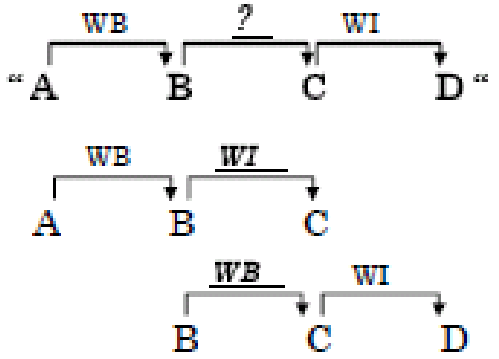
# 文字単位の係り受けによる表現(2)

- 縮退が起きているときは次のようにラベルを決める
- 構成要素の同じ部分の文字が縮退する時 → P
- 1つ目の構成要素の最後と2つ目の構成要素の最初の文字が縮退する時 → D
- この表現方法では1回の縮退で消える文字は必ず1文字でなければならない
- 2字以上の縮退は起こらないと思われるので問題なし

(a) “C”が縮退する場合（ABC+CD）




(b) “BC”が縮退する場合（ABC+BCD）



# 文字単位の係り受けによる表現(3)

- 並列かつ最初の文字が縮退するもの
  → ex.「国内外 (=国内+国外)」
  「国内+国外」とすると「国」の係り先が2つになってしまう
  →「国 => (内+外)」(先に「内外」を並列関係で結び、そこへ「国」が係る)と表現することで解決

- このような表現で・・・
  - 文字の縮退を自然に表現することができる
  - 従来の形態素解析と係り受け解析に相当する処理を1つの処理にまとめることができる

# 実験

- データ:ライフサイエンス辞書(LSD) 2008年4月版
  日本語94707語、英語83956語の専門用語を掲載
- このうち疾患を表す八文字以上の日本語から251語に対してタグ付けを行った
- タグ付けした語の構成要素の内部構造も同時に保存する
  ex.「急性骨髄性白血病」
  → 「骨髄性白血病」、「骨髄性」、「白血病」も同時に内部構造が定義される
  → タグ付けされた語は疾患名の他に解剖部位名や物質名を含む
- これらをあわせるとタグ付けされた語は計794語

# 実験(2)

- 文字対に対して係り受け関係の有無を識別するためTinySVMを用いて学習を行った
- カーネルは線形、使用した素性は以下のとおり

[文字情報]
- 係り元/係り先候補の漢字
- 係り元/係り先候補の文字種(ひらがな、カタカナ、漢字)
- 係り先候補の一文字後がカタカナかどうか

[文脈情報]
- 係り元と係り先候補の文字の距離
- 元の文字列中での位置(先頭、途中、最後)

[辞書情報]
- 係り元の文字と係り先候補の二文字から成る語が辞書に掲載されているかどうか
- 係り元/係り先候補で終わる語で、かつ元の文字列中に含まれる語が辞書に掲載されているかどうか

# 実験(3)

- この識別器を用い、係り先が1つ、交差は起こらないという制約を入れ、前方から順に係り先を決定する実験を行った
- 係り先はSVMの出力数値が最も大きなものを選択した

| Data | 文字対 | 語 |
|---|---|---|
| Baseline | 86.9% | 28.6% |
| System | 93.7% | 55.6% |

・ベースラインとして、全ての文字が次の文字に係るとしたときの値を示している

・システムの値は10分割交差検定によるもの

# 考察

- 構成要素間の係り受けが複数考えられることもある
  → 専門家であっても内部構造の定義は簡単ではない
- 作業者が考えるその言葉の意味を係り受け関係へ変換する
  → 構文木や係り受け関係の考え方を理解している必要あり
  → トップダウンに語を分割する方法は直観的でない場合がある
  ex.「甲状腺機能亢進」
    → 甲状腺(の)機能(が)亢進(する)
    →「甲状腺機能+亢進」と分割すべき
    しかし・・・「甲状腺(の)機能亢進」と捉えてそこで分割して
    しまいがち

# 考察(2)

- トップダウンな方法
  - 利点：構成要素として出現した語が既に内部構造を定義されていた場合に同じ定義を繰り返さなくていい

  → この利点を生かしたままボトムアップ的に作業できる環境が望ましい

- 実験結果より・・・
  - ベースラインは大幅に上回っているが実用には耐えない精度
  - 今回は共起情報など語についての素性を利用していない
  - LSD掲載語の一部には意味カテゴリが付与されている

    → これを取り入れられるように工夫したい

# おわりに

- 専門用語の内部構造の表現方法として以下のものを提案した
  - ラベル付きの構文木
  - 文字単位の係り受け
- SVMを用いて文字単位の係り受け解析を試みた
- 対象としたのは疾患名を始めとした生命科学分野の語
- 文字対に対する係り受け関係の有無の判定精度は93.7%
- 内部構造解析の精度は55.6%

  → 充分な精度ではない
- 今後の予定
  - 識別器の精度向上と内部構造解析のアルゴリズムの改良

    (他の手法も視野に入れる)