# 生命科学知識を得るために入力される自然言語クエリの構文解析

原 忠義[1]　建石 由佳[2]　Jin-Dong Kim[3]　宮尾 祐介[1]　[1] 国立情報学研究所　[2] 工学院大学　[3] ライフサイエンス統合データベースセンター

## 研究の目的

- 近年、多くの生命科学データベースが相互接続・統合されてきている
- そのため、高い表現力を持つクエリ言語 (例：SPARQL等) の必要性が増しているが、人が学んで用いるには複雑

→ 自然言語の文をそのまま入力クエリとして利用できるようにしたい

自然言語クエリ → **構文解析** → 解析木 → 変換 → SPARQL クエリ

*疑問:* ***最先端の構文解析技術およびその分野適応技術は自然言語クエリを正しく構文解析できるのか？***

## Step 1　自然言語クエリ文に木構造をアノテーション（ツリーバンクの作成）

### 対象とするクエリ文

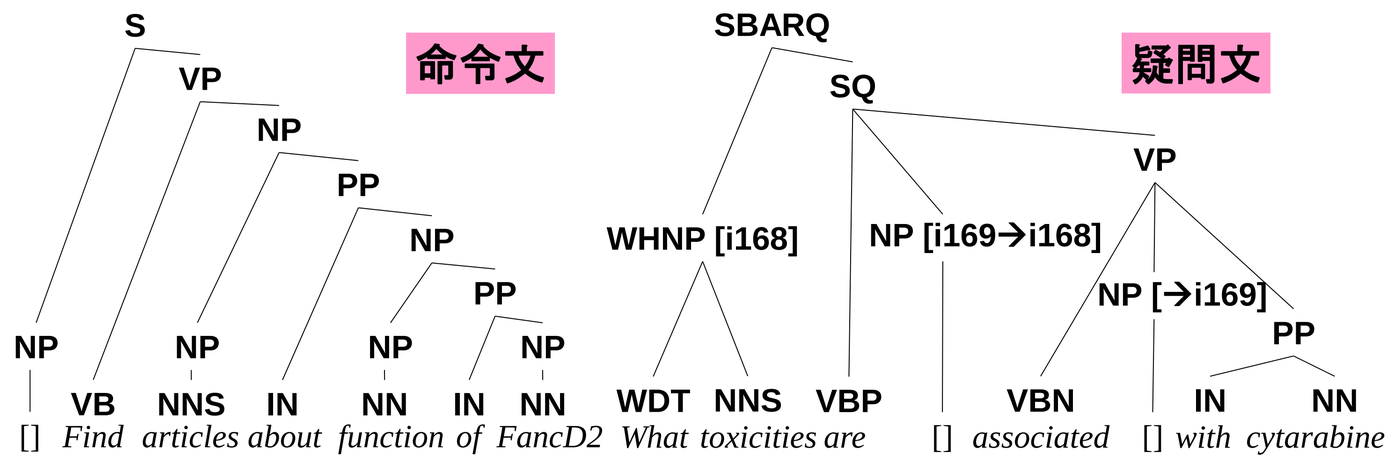
Genomics track (GTREC) '04-'07 [1] のクエリ文

| 文構造 | | 2004 | 2005 | 2006 | 2007 | 合計 |
|---|---|---|---|---|---|---|
| 命令文 | | 22 | 60 | 0 | 0 | 82 |
| 疑問文 | 疑問詞が名詞句 | 3 | 0 | 6 | 0 | 9 |
| | 疑問詞が修飾語 | 1 | 0 | 22 | 0 | 23 |
| | 疑問詞が限定詞 | 11 | 0 | 0 | 50 | 61 |
| | 疑問詞以外 | 5 | 0 | 0 | 0 | 5 |
| 名詞句 | | 14 | 0 | 0 | 0 | 14 |
| その他 | | 2 | 0 | 0 | 0 | 2 |
| 合計 | | 58 | 60 | 28 | 50 | 196 |

### 木構造アノテーション

GENIA Treebank (GTB) [2] のガイドランに従う

- 命令文 82 文 / 疑問文 98 文 / その他 16 文




## Step 2　アノテーションしたクエリ文上で品詞タガーと構文解析器の性能を調査

### 実験設定

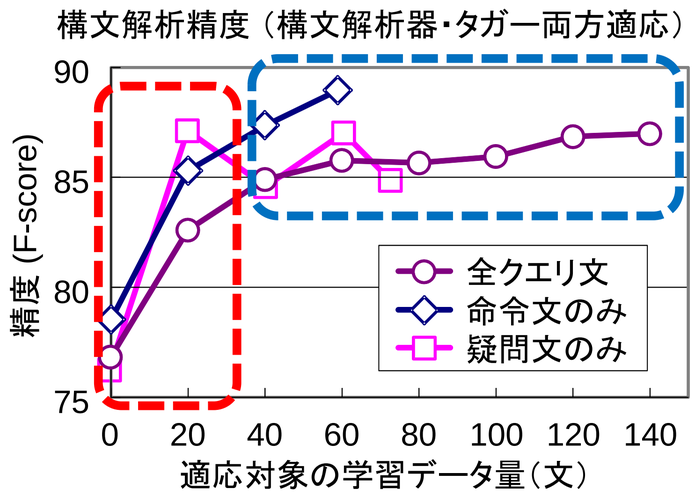
品詞タガー [3] ＋HPSG構文解析器 [4]
以下のデータ上で学習

- Penn ツリーバンク [5] 39,832 文
- GENIA ツリーバンク 14,849 文

➡ クエリ文ツリーバンクに適応

- 全クエリ文・命令文のみ・疑問文のみ
- 構文解析器 → 手法 [4] による適応
- 品詞タガー →
  PTB＋GENIA＋クエリ文で再学習

### 構文解析精度の概観

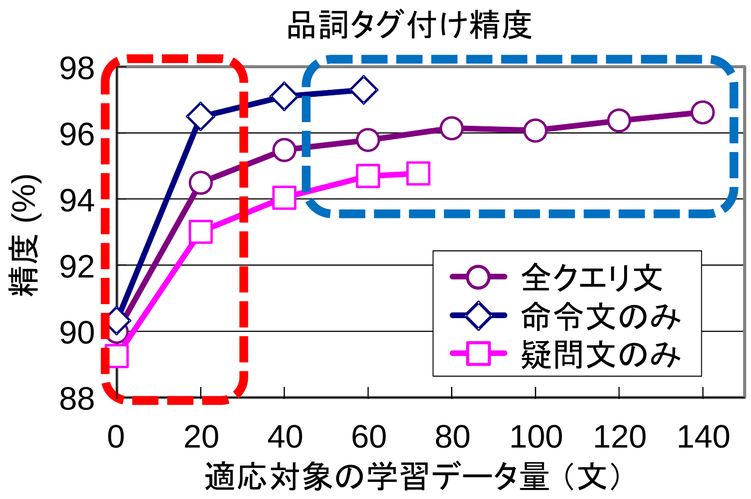


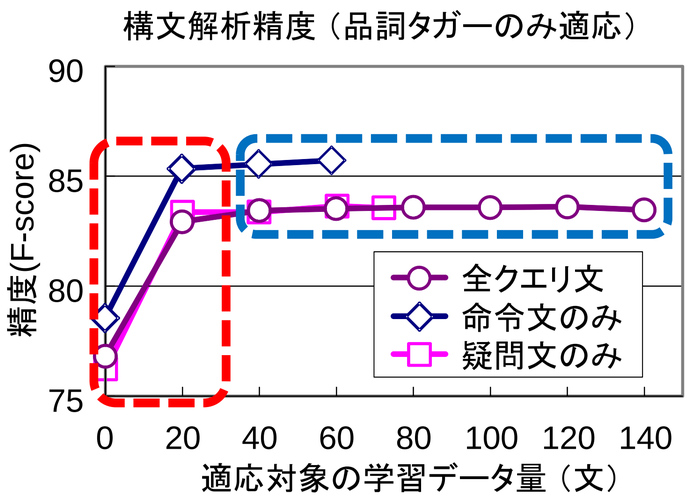

- 少量のツリーバンクで精度は著しく向上
- 改善傾向は持続しない

初期の飛躍的な精度向上　&
その後の改善傾向の減衰

➡ 更なる改善へ向けて、原因を調査

### 品詞タガーの性能




**改善されたエラー**　([ *正解* → *誤り* ])

- ***Find*** [VB→NN] *articles about ... .*
→ 構文解析時の「主動詞の欠落」
- ***What*** [WDT→WP] *toxicities are ... ?*
→ 構文解析時の「誤った疑問詞の振る舞い」

**クエリ文は主動詞等の多様性が低い**

➡ 少量のデータで容易に学習可能

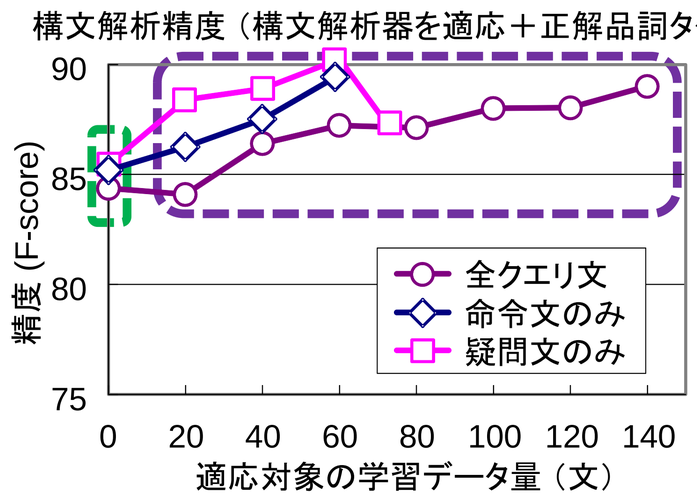
**適応後も残るエラー**　([ *正解* → *誤り* ])

- ***Mad***[JJ→NNP] ***Cow***[NN→NNP] ***Disease***[NN→NNP]
← アノテーションスタイルの差が原因
- *How do mutations ...* ***affect*** [VB→VBP] *...disorders?*
← タガーが長距離の依存関係を把握できていない

**学習データを増やすだけでは不十分**

➡ システムの再構築 ／
モデルの素性デザイン再考の必要性

### 構文解析器の性能




- 文の主構造に関わるエラーは観察されず

➡ 基本的にはクエリ文を解析できる

- より局所的な構造エラーが解決
- エラーのタイプによってはほとんど未解決
  - 等位構造 ／ 2つの節の接続

➡ 学習データを増やす／
モデル設計や枠組を再考する必要性

**参考文献**

[1] W. R. Hersh et al. 2004-2007. TREC 2004-2007 Genomics Track Overview. In Proceedings of the 13th-16th Text REtrieval Conference.
[2] Y. Tateishi et al. 2006. GENIA Annotation Guidelines for Treebanking. Technical Report, Tsujii Laboratory, University of Tokyo.
[3] Y. Tsuruoka et al. 2005. Developing a robust part-of-speech tagger for biomedical text. In Proceedings of 10th PCI.
[4] T. Hara et al. 2007. Evaluating impact of re-training a lexical disambiguation model on domain adaptation of an HPSG parser. In Proceeding of IWPT2007.
[5] M. Marcus et al. 1994. The Penn Treebank: Annotating predicate argument structure. In Proceedings of ARPA HLT Workshop.