社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－NL－159 (4)
2004／1／13

構文情報に基づく情報抽出システム開発のためのツール

保坂順子 Igor Kurochkin 小長谷明彦

理化学研究所 ゲノム情報科学研究グループ

〒230-0045 横浜市鶴見区末広町 1-7-22

{jhosaka, igork, konagaya}@gsc.riken.jp

生物・医学分野では、学術論文数の急増に伴い、文献からの情報抽出などの自動処理化が試みられている。しかし、抽出対象が多様であり、抽出システムを開発するための基礎データが不足しているという問題がある。我々は、構文情報を使った情報抽出システムの開発に向けて、基礎データ作成のためのツールキットを開発している。2種類の構文木と抽出を比較・編集することができ、抽出部分の変更を構文木に連動表示、編集結果をデータベースに保存することなどができる。

PBIE: A toolkit for developing parsing-based information extraction

Junko Hosaka, Igor Kurochkin, Akihiko Konagaya

Riken Bioinformatics Group

Suehiro-cho 1-7-22, Tsurumi-ku, Yokohama, Kanagawa, 230-0045 Japan

{jhosaka, igork, konagaya}@gsc.riken.jp

With the increasing amount of scientific literature in the biomedical domain, research can be accelerated by applying text processing approach including information extraction. However, the approach suffers from insufficient basic data. We extract information based on parsing output. To prepare the basic data, we are developing a toolkit that combines annotation tool and tree editing tool. Any change in annotation can be simultaneously viewed on the syntactic tree. The result can be stored into a database using a conversion tool. Comparing two sets of extraction, the performance can be automatically calculated.

1. はじめに

生物学・医学の分野では、近年文献数が膨大になり、その自動処理化が不可欠になってきている。たとえば、免疫機能の活性化に寄与すると言われている、インターロイキン6に関する文献を、"Interleukin 6" をキーワードとして PubMed[1]で検索すると、2003年1月1日から12月までで1730件、期限を設けないと12月検索時点では22986件の文献が検索される。

文献の自動処理化を目指し、たんぱく質間相互作用抽出に代表される、生物学・医学文献からの情報抽出が盛んに行われている。単語の共起を使ったもの [1]、フルパーザを使ったもの [2]、抽出規則を人手で書き下したもの [3]、医学文献用に開発したパーザを、分子生物学用に変更を加えたもの [4] などがある。しかし、ある程度実用化できる抽出システムを開発するためには、基礎データが不足している。

我々は、抽出情報の多様さに対処するため、構文解析パーザを使い、その結果を基に抽出を行っている。その基礎データ作成のために、生物学者と言語学の専門家の要

[1] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

求を反映したツールキット、Parsing-based Information Extraction Toolkit, PBIE を開発している。本稿では、抽出部分の比較・編集、構文解析結果の比較・編集、抽出編集と連動した構文木上の表示などをするためのツールについて報告する。

PBIE ツールキットは Windows2000/XP の環境で動作確認しており、Microsoft Internet Explorer 6.0 以上が必要である。また、データの保存には Microsoft ACCESS を使っている。

2. PBIE ツールキットの概要

開発中の PBIE ツールキットは、以下の 5 点の容易さを考慮してデザインした：

- 抽出表示色の変更
- 抽出カテゴリーの変更
- 複数の構文パーザ結果を使った実験
- 抽出箇所と構文木の関係分析
- コメントの集計

この PBIE ツールキットは、PBIE 評価ツール(PBIE)と、XST データ管理ツール(XST)からなっている。

2.1. PBIE 評価ツール

PBIE は、実行ファイルと、3 つの XML 形式のファイルで構成されている。後者は、差し替えが可能である。ファイル構成を**表 1**にしめす：

ファイル名	内容
PBIE.exe	実行ファイル
category.xml	抽出カテゴリーのリスト スタート・モードの定義
nodenames.xml	品詞・構文ノードのリスト
comment.xml	コメントのリスト

表 1: PBIE 評価ツールのファイル構成

PBIE では、2 種類の構文解析結果と情報抽出結果が比較評価できる。たとえば、パージング結果と、それに人手で修正を加えたものの比較や、同一パーザで辞書を入れ替えて解析した結果などを比較することが考えられる。情報抽出も、自動的に抽出したものと、人手で抽出したものを比較したり、生物学者と言語学者がそれぞれ抽出箇所をマーキングしたものを比較することなどが考えられる。この評価ツールでは、前もって解析、抽出したものを表示して比較するだけでなく、編集もできる。特に、抽出箇所をマーキングすると、それと連動して、構文木の対応箇所が同じ色で自動的にマーキングされる。これは、構文解析結果を利用した情報抽出規則を作成するのに役立つと考える。

現在使っている抽出カテゴリーのリストとコメントのリストは、生物学者一人と言語解析の専門家二人がたんぱく質間相互作用に関する文を評価した際のコメントを参考にして作成した。前者は、400 文の評価を行い、後者は、さらに 600 文、合計 1000 文の評価を行った。

2.2. XST データ管理ツール

XST は、外部からの入力、PBIE、データベースの 3 つがそれぞれ同じデータを処理できるようにデータ形式を変換するツールである。入力文やそのパージング結果を取込む際に、これらを評価ツールで処理できる形式に変換したり、データベースに保存してあるものを組み合わせて評価ツールで処理できる形式に変換したりできる。また、評価ツールで評価した結果をデータベースに保存する形に変換できる。

図 1に、XST のユーザ・インターフェイスを示す：

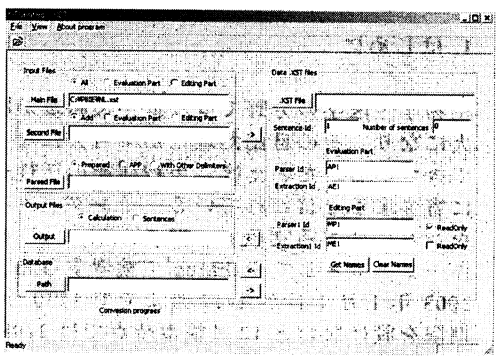

図 1：XST のインターフェイス

図 2に、主に XST を介した、入・出力、PBIE およびデータベース間のデータの流れを示す：

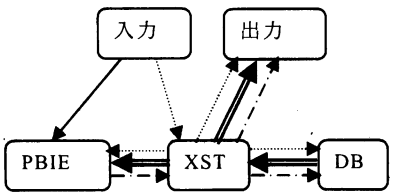

図 2：XST とデータの流れ

入力としては、以下の 3 種類のものを受ける：
1．テキスト形式の文のリスト
2．パーザの出力
3．PBIE で処理可能なファイル

このうち、1 と 2 は図 2では、細かい点線で表してあり、XST で PBIE 用に変換する。2 は、ニューヨーク大学で開発された Apple Pie Parser[2] (APP)とシュトゥットガルト大学で開発された LoPar[3] で変換の確認をした。3 は、図 2では、実線で表してあり、直接 PBIE に取込める。PBIE で評価したものは、XST で変換して、出力またはデータベース化できる。同一文について、2 種類の構文解析、情報抽出およびコメントの PBIE 入力構造を、次に示す。

"parsed", "extracted", " commentp" が、一番目のセットで、"parsed1", "extracted1", "commente1" が、二番目のセットである：

```
<sentence id=  sentenceid= >
<original> </original>
<parsed></parsed>
<extracted>
        <pharse start= end=> </phrase>
</extracted>
<parsed1 sign =  > <parsed1>
<extracted1 sign =  >
        <phrase start= end=> </phrase>
</extracted1>
<commentp> </commentp>
<commente> </commente>
<commentp1> </commentp1>
<commente1> </commente1>
</sentence>
```

例として、1 種類の評価の一部を次に示す：

```
<sentence id="8" sentenceid="108">
<original>PAK I is activated by Cdc42 ( GTP ) .
</original>
<parsed><![CDATA[<S><S><NP><NPL><NNPX>PAK
</NNPX></NPL><NPL><PRP>I</PRP></NPL></NP><
VP><VBZ>is</VBZ><VP><VBN>activated</VBN>.....
</VP></VP></S><_PERIOD>.</_PERIOD></S>]]></pa
rsed>
<extracted patternid="1">
<phrase type="recipient" start="0" end="4">PAK I
</phrase>
<phrase type="verbal" start="6" end="17">is activated
</phrase>
<phrase type="agent" start="19" end="34">by Cdc42
( GTP ) </phrase>
</extracted>
<commentp/><commente/></sentence>
```

出力形式は、2 種類ある：
1．テキスト形式の文のリスト
2．精度計算

精度計算は、抽出部分に対して行われる。これは、2 種類の抽出結果を比較し、一方を正解とみなし、再現率、適合率および F 値を計算する。

データベースに保存してあるものは、異なる評価を組み合わせて XST で変換し、PBIE で再度評価したり出力したりできる。

2.3. スタートアップ

PBIE で使用できる抽出部分のマーキングと構文木の編集の 6 種類の組み合わせを表 2に示す。2 種類のマーキング、および構文解析木が同時に見られる場合を、"evaluation"としている：

[2] http://www.cs.nyu.edu/cs/projects/proteus/app/

[3] http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/LoPar-en.html

	組合せツール
1	Extraction Marking
2	Extraction Evaluation
3	Sentence Tree Editing
4	Sentence Tree Evaluation
5	Extraction Marking and Sentence Tree Editing
6	Extraction and Sentence Tree Evaluation

表 2：PBIE スタートアップメニュー

どの組合せツールで開始するかは、category.xml で定義する。さらに、抽出カテゴリーで使っている色の編集を可・不可にするかもこのファイルで定義する。

組合せツール番号６で、色の編集を可能にした場合の画面を図 3に示す：

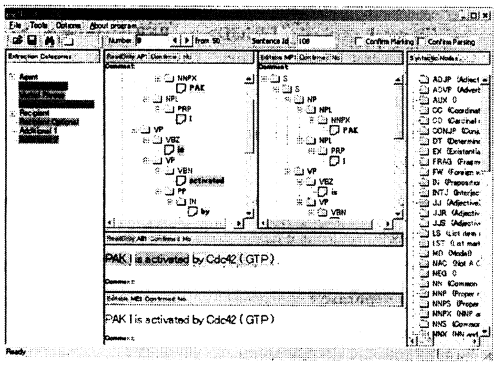

図 3：PBIE Extraction and Sentence Tree Evaluation インターフェイス

2.4. 抽出カテゴリー

抽出に使用するカテゴリーは、Agent, Verbal Phrase, Recipient など 8 つ用意してある。そのうち 2 つについてはさらに、Agent Compound, Agent Process などサブカテゴリーを用意した。これは、生物学者が 400 文について評価したときの経験を基に作成した。

PBIE では背景と文字の色が変更でき、この変更を保存することもできる。

2.5. 木構造のノード

APP で使用している 70 のノードを前もって定義しているが、これはほぼ Penn Tree Bank で定義されているものと同じである。このうち、次の 4 ノードが APP 固有のものである：SS(サブＳ), NPL(最下位の名詞句), NNPX(固有名詞の単複), NNX(名詞の単複)。

2.6. コメント

コメントは属性と属性値という形で記入できる。値は数値に限らず、アスキー文字を使って自由に記入できる。また、コメント用の属性は comment.xml ファイルを編集することにより書き換えられる。

構文解析に対するコメントには、８つの属性が用意してある。そのうち７つは、言語解析の専門家二人が、たんぱく質間の相互作用に関する文を評価した際のコメントを参考にしたものである。これ以外のものは、Other comment として記入できる。

Whole sentence analysis
Coordination
Assigning Part of Speech
Ordinary relative clause
Reduced relative clause
Subordinate clause
Noun phrase identification
Other comment

情報抽出に対するコメントには、５つの属性が用意してある。そのうち４つは、生物学者一人が、自動抽出を評価した際のコメントを参考にしたものである。構文解析部と同様、これらの属性で扱いきれないものは、Other comment として記入できる。

Main verb
Acting element
Receiving element
Act-Receiving element
Other comment

3. PBIE を使った編集

PBIE では、木構造および抽出文字列選定の編集ができる。また、文字列による文検索もできる。

3.1. 木構造の編集

マウスボタン操作により、ノードを挿入・削除したり、移動、名前の変更などができる。また、前もって定義されたノード名をドラッグ・アンド・ドロップで使うこともできる。さらに、編集した木構造は、保存して再利用できる。

図 4に、木構造編集時のインターフェイスを示す：

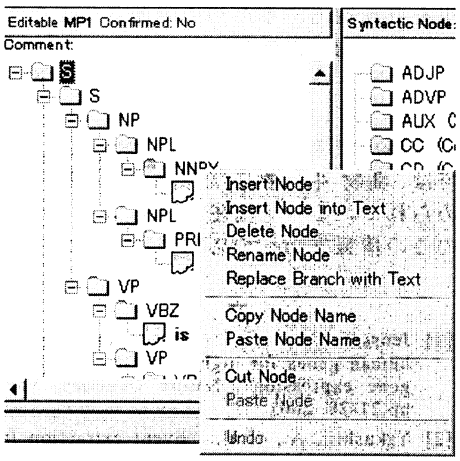

図 4：木構造の編集

3.2. 抽出マーキング

マウスで範囲指定し、カテゴリーをクリックすることにより、マーキングできる。範囲指定が単語の中間になった場合は、自動的に単語の境界までマーキングされる。単語の境界は、空白にしている。マーキングするのと同時に構文木の該当箇所も同色でマーキングされる。図 5に、抽出箇所をマーキングするのに連動して、文字列が構文解析木に同色で表示される例を示す。該当箇所はマルで囲んである：

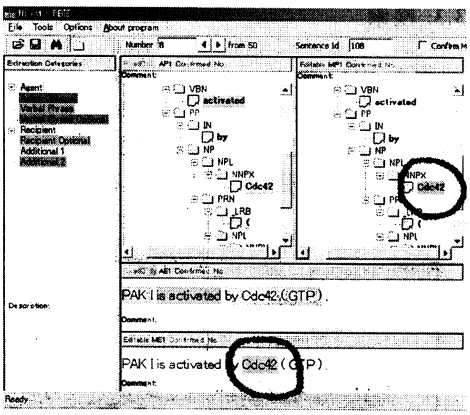

図 5：抽出編集と構文木上の表示の連動

抽出のマーキングに使う色およびカテゴリーを表す文字の色は、PBIE 上で使用者が適宜変更することができる。マウスボタン操作により、色の選択モードに入る画面と、マーキングまたは文字の色の設定を選択したときに使用できる色の種類を、それぞれ図 6図 7に示す：

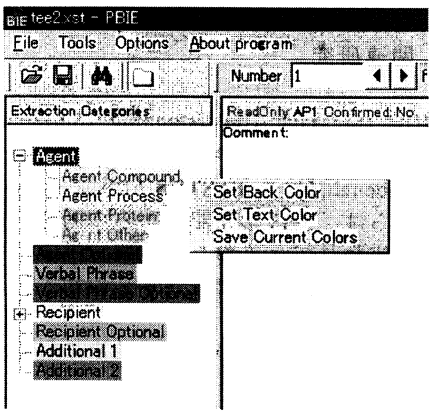

図 6：抽出カテゴリー色の設定

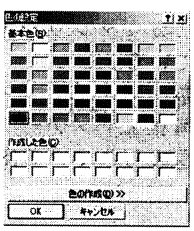

図 7：抽出カテゴリーに使用できる

3.3. 検索

文番号を指定して、その文に移動することもできるが、文中に現れる文字列を指定して検索することもできる。図 8に、文字列"Cdc42"を指定して検索した結果を示す：

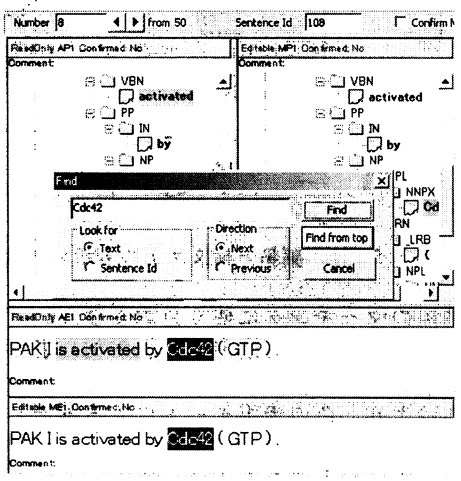

図 8：文字列検索

4. 関連研究

MUC (Message Understanding Conferences) では、情報抽出のツールとして、Alembic Workbench[5] や FASTUS[6]が開発された。これらは、主に新聞記事からの固有名詞抽出に使われてきており、生物・医学分野での応用報告は、本稿報告者らの知る限り、行われていない。

生物・医学分野の項情報抽出の経験から開発されているツールとして、willex がある[7]。これは、汎用的だといわれている文法を、生物・医学分野の文の解析に使えるように、改良をするためのデバッグツールであり、特にカヴァレッジをあげることに重点を置いている。

また、医学に関するテキストを解析し、さらに言語情報をアノテーションするツールの開発も進められている[8]。

5. おわりに

本稿で紹介した PBIE ツールキットは、基礎データ作りに有効利用できると考える。しかし、ツール上で規則を構築し、自動抽出することはできない。今後は、パーザや情報抽出モジュールを組込み、一連の作業を簡略化する予定である。

構文解析に基づく情報抽出のための PBIE ツールキットは、たんぱく質の相互作用を自動抽出するために開発されてきた。しかし、抽出カテゴリーや前もって定義したコメントなどを変更することが容易だというように、自由度が高いため、他の分野でも利用できると考える。たとえば、薬学では、薬物-生体物質関係などの研究もすすめられている[9]。現在、PBIE ツールキットのこの情報抽出への応用が計画されている。

文　献

[1] Jenssen, T-K., et al.: "A literature network of human genes for high-throughput analysis of gene expression", Nature Genetics, Vol.28, pp.21-28, 2001

[2] Yakushiji, A., et al.: "Event extraction from biomedical papers using a full parser", Proc. of PSB-2001, Vol.6, pp.408-419, 2001

[3] Blaschke, C. and Valencia, A.: "The potential use of SUISEKI as a protein interaction discovery tool", Genome Informatics, Vol.12, pp.123-134, 2001

[4] Friedman, C., et.al.: "GENIES: a natural –language processing system for the extraction of molecular pathways from journal articles", Proc. of ISMB-2001, Vol.17 Suppl.1, pp.S74-S82, 2001

[5] Aberdeen,J., et al.: "MITRE:Description of the Alembic system as used in MET", Proc. of the TIPSTER 24-Month Workshop, pp.461-462, 1996

[6] Appelt,D.J., et al.: "SRE international FASTUS system MUC-6 test results and analysis", Proc. of the Sixth Message Understanding Conference, pp.237-248, 1995

[7] 薬師寺あかね et al.: "実用的な文法を開発するためのデバッグツール", 情報処理学会研究報告 NL-155, pp.19-24, 2003

[8] Grover, C., et al.: "XML-based NLP tools for analysing and annotating medical language", Proc. of the 2nd Workshop on NLP and XML, 2002

[9] 吉川澄美, 小長谷明彦: "薬物と生体物質の相互作用オントロジーに基づく薬機能知識ベースの設計", 臨床評価, Vol.29, No.2・3, pp.275-286, 2002