PBIE2: 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　　松村和美　　吉川澄美　　Igor V.Kurochkin　　小長谷明彦

理化学研究所　ゲノム科学総合研究センター
横浜市鶴見区末広町 1-7-22
{jhosaka, kmkazumi,sumi,igork,konagaya@gsc.riken.jp}

要旨

構文情報に基づく情報抽出をめざして、パージング、抽出規則の作成、抽出句のマーキングなどをダイナミックに行うツール PBIE２を開発した。我々は、薬学や生物学などの学術文献からの薬物‐たんぱく質間相互作用などの情報抽出を試みている。一般的な構文解析パーザを使い、その解析結果に抽出規則を適用して、情報を抽出するという手法をとっている。しかし、辞書に専門用語を追加しただけでは、解析精度は不十分だと想定されるため、構文規則の改良が必要である。また、抽出規則を作成するには、抽出したい情報と構文との関連を考慮する必要がある。PBIE２は、この一連の作業を効率的に行うためのツールである。

1　はじめに

生物学・医学の分野では、近年学術文献数が膨大になり、その自動処理化が不可欠になった。

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。

2001 年に開始された GENIA プロジェクト[1]では、分子生物学分野の学術論文の抄録にたんぱく質、DNA,、RNA などの情報を付与したコーパスを作成している。2003 年 3 月にはそのサブセットの 2,000 本の抄録に形態素情報を付与したコーパスもリリースされた。また、BioCreative[2]では、コンテスト形式で、ヒトゲノム関連の学術論文の本文に Gene Ontology[3]で定義された機能情報を付与し、その正解コーパスは、専門家の協力を得て作成している。このように、共通に使えるコーパス作成は進んでいるが、その作成には、多大な労力を要する。さらに、コーパスは、分野ごとに必要である。

我々は、少ないデータで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求を反映したツールキット、Parsing-based Information Extraction Toolkit を開発している。PBIE [5]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実装した。これは、主に、分野の専門家が抽出箇所をマーキングしたものを正解として、言語の専門家が自動的に抽出したものと比較検討し、基礎データを作成することを目的としている。PBIE ではパージングと抽出は別途行っていたが、作業の効率化のため、PBIE２では、パーザのプラグイン機能と抽出規則作成の機能を追加した。PBIE２で使用するデータは、XML 形式をとっており、抽出規則作成には、xPath の記述を使っている。また、自動抽出の精度を計算、データの組合せを変更、データベースに格納するための変換などを行うツールも、合わせて開発した。なお、本稿では、ApplePie Parser ver.5.9[4](APP)をプラグインしてパージングした結果を使っている。

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

[2] http://www.pdg.cnb.uam.es/BioLink/BioCreative.eval.html

[3] http://www.geneontology.org/

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

PBIE２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　PBIE２の概要

PBIE２では、マーキング用のウィンドウと構文解析木用のウィンドウが、それぞれ最大２つまで表示できる。これは、各分野の専門家が抽出箇所をマーキングしたものと言語の専門家が自動的に抽出したものを比較検討したり、2 種類のパージング結果や 2 種類の人手による解析結果を使ったりするためである。PBIE２では、さらに、プラグインしたパーザでパージングしたり、抽出規則を作成して、その場で自動抽出を実行したりすることもできる。

2.1　PBIE２の構成

PBIE２は、実行ファイル、４つの XML 形式のファイル（.xml, .xpt）、および外部からパーザを取り込むためのファイル（.dll）で構成されている。実行ファイル以外は、差替えが可能であり、xpt ファイルはファイル名の変更も可能である。ファイル構成を表 1に示す：

ファイル名	内容
PBIE2.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
comment.xml	コメントのリスト
nodename.xml	品詞・構文ノードのリスト
Pattern.xpt	抽出規則のリスト
Parser.dll	パーザ取込み用ファイル

表 1: PBIE２のファイル構成

2.2　PBIE２の入力ファイル

入力として、テキストファイル（.txt）と xst ファイルを受ける。テキストファイルは、一文一行を前提として、PBIE２で xst の形式に変換し、パージングまたはマーキングに使う。保存は、テキスト形式など指定できるが、デフォルトでは xst ファイルになる。これは、入力としてそのまま再利用できる。

PBIE２が受ける xst 入力ファイルのデータ構造の一部を、次に示す。original に囲まれる文について、2 種類の構文解析、情報抽出およびコメントの記入ができる：

<sentence id = *s_id* **sentenceid** = *sent_id* **>**
<original> *original_sentence* **</original>**
　　<parsed> *parsed_sentence* **</parsed>**
　　<extracted>
　　　　<phrase type = *ph_type* **start** = *s_no*
　　　　end = *e_no***>**
　　　　extracted_phrase **</phrase>**
　　</extracted>
　　<parsed1 sign = *pconf_on/off* **>**
　　parsed_sentence1 **<parsed1>**
　　<extracted1 sign = *econf_on/off* **>**
　　　　<phrase type = *ph_type* **start**= *s1_no*
　　　　end= *e1_no***>**
　　　　extracted_phrase1 **</phrase>**
　　</extracted1>
　　<commentp> *com_parsing* **</commentp>**
　　<commente> *com_extraction* **</commente>**
　　<commentp1> *com_parsing1* **</commentp1>**
　　<commente1> *com_extraction1* **</commente1>**
</sentence>

parsed, extracted, commentp, commente が、一番目のセットで、 parsed1, extracted1, commentp1, commente1 が、二番目のセットである。 start, end は先頭からの文字の位置を、sign はユーザが解析、または抽出を確認したかどうかを示す。

2.3　PBIE２の起動

PBIE２では、抽出部分のマーキング、構文木の編集、抽出規則の編集が７種類の組合わせで使える。これらの組合せは、スタートアップメニューで選択できる。組合せを表 2に示す。2 種類のマーキング、および構文解析木を同時に表示する場合を、Evaluation としている：

番号	組合せツール
1	Extraction Marking Editor
2	Extraction Evaluation
3	Sentence Tree Editor
4	Sentence Tree Evaluation
5	Extraction Marking and Sentence Tree Editor
6	Extraction and Sentence Tree Evaluation
7	Extraction Pattern Editor , Extraction Evaluation and Sentence Tree Editor

表 2：PBIE２スタートアップメニュー

表 2に示すように、抽出規則の編集をする場合は、ツール番号 7 に示す組合せに限っている。これは、抽出規則作成は、解析結果を基に行うために解析木の参照が必要であり、編集した規

則の有効性を調べるには、正解例または変更前の抽出結果と比較することが必要だからである。

PBIE 2 の開始ツール、使用可能ツールを設定するモード、スタートアップメニューの表示の有無、起動時に表示するツールの番号、抽出カテゴリーで使っている色の編集を可・不可にするかは、category.xml で定義する。これらの定義例を、抽出カテゴリーAgent の定義例と共に示す。7 種類の組合せが使えるモード、スタートアップメニュー非表示、開始ツール番号 7、抽出カテゴリーの色の編集可に設定されている。bckcolor と txtcolor の値は、RGB 値である：

<?xml version="1.0"?>
<categories mode="7" startup="0" tool="7" coloreditor="1">
<agent id="1" name="Agent" category="agent" bckcolor="16710867" txtcolor="0">
Agent
Example:
We find that ACK-2 can be activated by cell adhesion in a Cdc42-dependent manner.
Agent: cell adhesion

以上の設定で PBIE 2 を起動し、抽出カテゴリーの Agent を選択したところを図 1に示す：

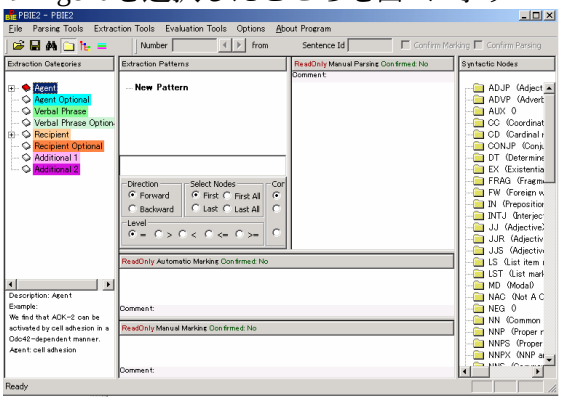

図 1： PBIE 2 起動直後、Agent 選択時の画面

抽出カテゴリーは、生物学者がたんぱく質間相互作用に関する 400 文を評価した際のコメントを参考にして作成した。Syntactic Nodes では APP の 70 のノードを定義してあり、これはほぼ Penn Tree Bank のものと同じである。

3 パージング

PBIE 2 で文をパーズするには、まず文のリストを取込み、プラグインしたパーザにかける。ツール番号 5 で、文のリストを取り込んだところを図 2に示す：

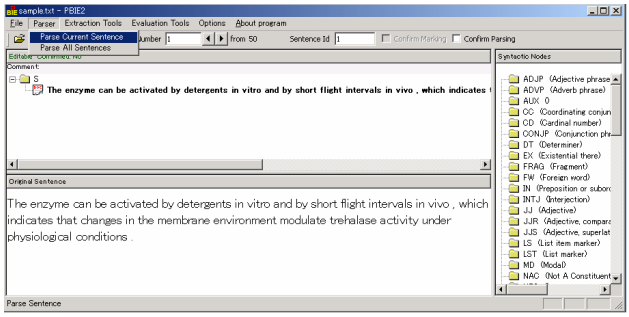

図 2: 文の取込み

パージングは一文ごとに実行するか、取込んだ文集合全体に実行するか選択できる。パーザを使って解析した結果は図 4のように表示される。

4 情報抽出

抽出規則の記述には、方向、レベルなどの情報の他に xPath [5]を使い、さらに詳細な指定をする。

4.1 抽出規則編集のインターフェイス

抽出規則の編集ウィンドウは 3 つの部分から構成されている。上部では、規則の構成を定義する。規則はひとつ以上のパターンの集合からなっていて、それぞれのパターンはステップからなっている。ステップは、終始ステップのような特殊なもの以外は、編集ウィンドウの左側にある Extraction Categories のカテゴリーをドラッグして作成する。コピー、貼り付け、削除などは、編集用のメニューをマウスボタンのクリックで呼び出して行う。編集メニューを呼び出したところを図 3に示す：

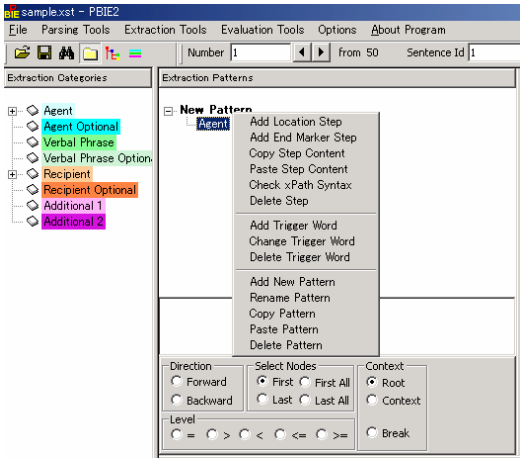

図 3: 抽出規則編集

[5] http://www.w3.org/TR/1999/REC-xpath-19991116

中央部では、xPath で抽出部分を詳細に指定する。下部では、方向、木構造上の階層などを、ボタン式で指定する。

4.2 抽出規則作成および実行

抽出規則の定義ファイル Pattern.xpt のステップに関するデータ構造を示す。name, type などの属性の記述は任意で、順番も自由である：

<step id= *stepid_no* **name =** *category_name* **type =** *category* **direction=** *direction* **level** = *level_mode* **select =** *selection* **no=** *break_mode* **>**
  <stepxpath id= *xpathid_no* **>**
    xPath expression
  </stepxpath>
</step>

例として、解析結果を基に作成した、動詞句と被動作主の抽出規則を示す。対応する PBIE 2 のインターフェイスは、図 4である：

```
<step id="3" name="Verbal Phrase" type="verbal" select="first" no="break">
        <stepxpath id="4">//VP[descendant::VBN ="activated"][VP or VBN or VBZ][not(descendant::SBAR//VBN)][not(CC)][1]
        </stepxpath>
</step>
<step id="6" name="Recipient" type="recipient" direction="forward" select="last_all" level="=" no="break">
        <stepxpath id="7">//NPL|//NP|//NNPX
        </stepxpath>
</step>
```

このようにして記述した動詞句、動作主、被動作主を抽出する規則を、パージング結果に適用した結果を図 4に示す：

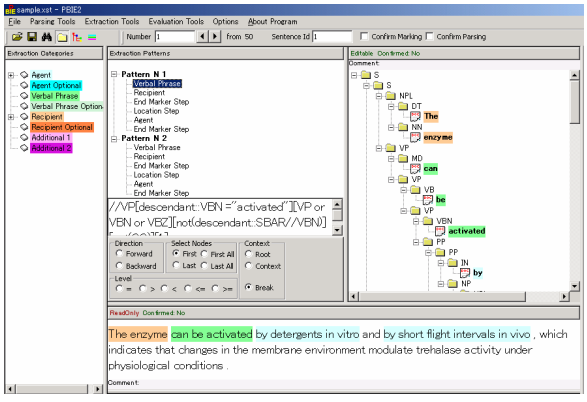

図 4: 抽出規則を使った情報抽出

5 関連研究

MUC (Message Understanding Conferences) では、情報抽出のツールとして、Alembic Workbench[6] や FASTUS[7]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[8]。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[9]。

6 おわりに

本稿で紹介した PBIE 2 は、構文解析に基づく情報抽出のための基礎データ作りに有効利用できると考える。今後は、検索機能の充実、抽出規則の自動生成などをめざす予定である。

また、薬物-生体物質関係などの研究 [10]に応用する予定である。

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] 保坂順子 et al.: “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”, 情報処理学会研究報告 NL-159, pp.19-24, 2004

[6] 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

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

[8] 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

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

[10] 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