# 述語項構造パターンを用いた医学・生物学分野情報抽出

薬師寺 あかね [1]　宮尾 祐介 [2]　建石 由佳 [3,1]　辻井 潤一 [2,3]
[1] 東京大学大学院情報理工学系研究科コンピュータ科学専攻
[2] 東京大学大学院情報学環　　[3]CREST, 科学技術振興機構
{akane, yusuke, yucca, tsujii}@is.s.u-tokyo.ac.jp

## 1　はじめに

医学・生物学分野において情報抽出システムを構築する際には、個々の細分野に対応するためのシステムの移植性が重要となる。我々は、より少ない正解付きコーパスから抽出ルール構築を行うため、構文解析の結果である述語項構造のパターンを用いる手法を提案する。我々の手法では、まず構文解析器で多様な統語上の変形を正規化した述語項構造を得る。その述語項構造の上で、表層上の単語列を用いるよりも一般化されたパターンを構築する。

近年、バイオインフォマティクスへの情報抽出技術の適用が多く試みられているが、実際に実用化されている例は少ない。これは、現在の情報抽出技術は精度が十分でないため、および各細分野への適用に多くの人手による労力を必要とするためである。人が書いた情報抽出ルールを用いた既存システム [1, 2] では、ほぼ実用に適した精度を得られているものの、これらのルールは他の細分野へ移植するのに多くの人手による修正が必要である。

一方、機械学習によって抽出ルールを構築するシステムでは、2 通りの問題がある。Huang らによるシステム [3] では、得られた抽出ルールは人にとって見やすいものであるものの、学習コーパスでは同一イベントを表す文のアライメントが与えられていなければならない。Bunescu らによるシステム [4] では、タンパク質間相互作用抽出ルールを学習するためのコーパス Aimed は相互作用の関係にあるタンパク質の語が指定されているだけである。しかし彼らの手法では、表層上の単語列に基づくパターンを獲得しているために得られるルールは十分に一般化されておらず、またさらに人手によって改良することが困難である。

より包括的な方法を取るためには、情報抽出システムを分野が異なっても共通の部分と各細分野に特化した部分とに分ける必要がある。そこで我々は、一般的言語知識を用いる構文解析器を組み込んだ情報抽出システムを構築した。構文解析器の出力は述語項構造であり、これを元に各細分野に対応した抽出ルールを次の 3 段階で自動的に獲得する：まず、求める情報を含む述語項構造上のパターンを抽出する（**パターン抽出**）。次に、パターンをさらに一般化するため部品に分割する（**パターン分割**）。最後に、学習コーパス全体と照らし合わせて実際には不適切であるようなパターンを削除する（**パターン選別**）。

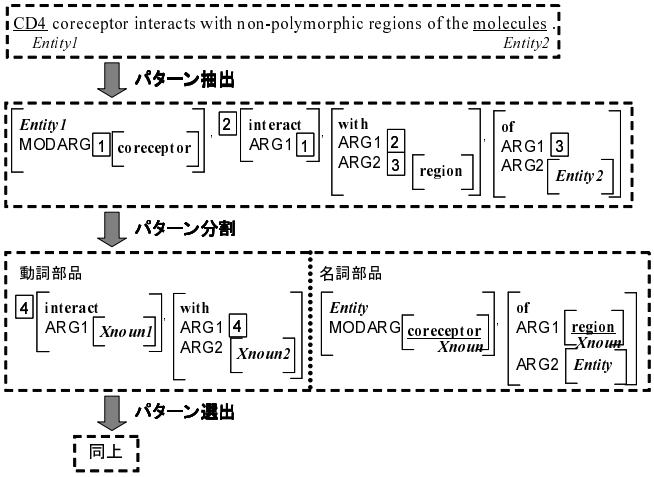



図 1: 抽出ルール構築の流れ

実際にタンパク質間相互作用の抽出実験を既存研究 [4] と同じ Aimed を用いて行った結果、適合率 75.6%、再現率 43.8%を得た。これは他の既存研究と比較しても十分に有効なものである。本稿では、今回取った手法および実験結果と分析を報告する。

## 2　手法

我々のシステムでは、相互作用の関係にあるタンパク質の組がタグ付けされた学習コーパスから、抽出ルールを自動構築する。その後、この構築された

| | |
|---|---|
| *Entity1* recognizes and **activates** *Entity2*. | |
| *Entity1* can **activate** *Entity2* through a region in its carboxy terminus. | |
| *Entity2* are **activated** by *Entity1a* and *Entity1b* | |
| *Entity2* **activated** by *Entity1* are not well characterized. | |
| The herpesvirus encodes a functional *Entity1* that **activates** human *Entity2*. | |
| *Entity1* can functionally cooperate to synergistically **activate** *Entity2*. | |
| The *Entity1* play key roles by **activating** *Entity2*. | |

**activate**
ARG1 *Entity1 (semantical subject)*
ARG2 *Entity2 (semantical object)*

図 2: “activate”の構文的変形とその述語項構造

抽出ルールを新たなテキストに適用し、相互作用しているタンパク質の組を抽出する。実際にコーパス中の文から抽出ルールを獲得した例を図 1[1]に示す。以下、構築の具体的手法について述べる。

## 2.1 構文解析

構文解析に用いるのは、Head-Driven Phrase Structure Grammer [5] に基づいた構文解析器 Enju [6] である。Enju は英語コーパス Penn Treebank [7] から文法および曖昧性解消モデルを学習した構文解析器である。

Enju による構文解析によって、例えば図 2 中の文における *Entity1* と *Entity2* の関係はすべて図中右の述語項構造で表される。

## 2.2 述語項構造パターンの抽出

構文解析の後、相互作用の関係にあるタンパク質の語をつなぐような最小の述語項構造の集合 $(p_0, p_1, \ldots, p_n)$ を抜き出しパターンとする。タンパク質が複数の語で表されている場合は、最後の語を代表語として扱う。パターン $(p_0, p_1, \ldots, p_n)$ を抜き出す実際の処理は以下の通りである：

$including(p)$ : $p$ の項あるいは修飾先であるような述語項構造の集合

$included(p)$ : $p$ を項あるいは修飾先として含むような述語項構造の集合

1. Enju によって文を述語項構造の集合に変換する。

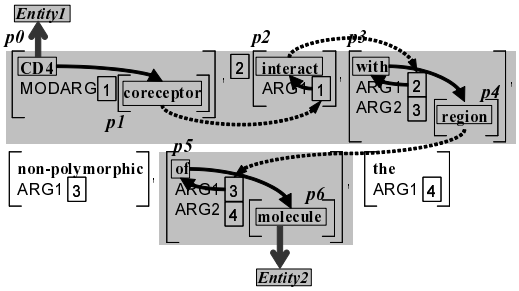
2. 相互作用の関係にあるタンパク質のうちの一方に対応する述語項構造を $p_0$ とし、$p_i = p_0$ として以下の 2-1 – 2-2 でパターン $(p_0, p_1, \ldots, p_n)$ の候補集合を求める：

   2-1, $p_i$ がもう一方のタンパク質なら、ここまでの $(p_0, \ldots, p_i)$ をパターンの候補とする。

   2-2. そうでなければ、$including(p_i)$ あるいは $included(p_i)$ に含まれる述語項構造それぞれを $p_{i+1}$ としてパターン候補を分岐させ、2-1 から繰り返す。

3. パターン候補のうちもっとも要素数が少ないものを選び、パターンとする。

4. パターン中の今回注目したタンパク質の語を変数に置き換える。




*CD4* $_{Entity1}$ **coreceptor interacts with** non-polymorphic **regions of** the *molecules* $_{Entity2}$.

図 3: 述語項構造パターンの抽出例

図 3 に例を示す。図中の文において “*CD4*” と “*molecules*” が相互作用にあるタンパク質の語とする。まず、“CD4” に対応する述語項構造を $p_0$ とする。$including(p_0)$ が “coreceptor” の述語項構造を含むため、それを $p_1$ とする（実線の矢印）。次に、$included(p_1)$ が “interact” の述語項構造を含んでいる（点線の矢印参照）ため、それを $p_2$ とする（次の実線の矢印）。同様にして “molecules” まで続けると、パターンとするべき述語項構造は図中影が付けられたもの $(p_0, \ldots, p_6)$ となる。最後に “CD4” と “molecules” をそれぞれ変数 *Entity1* と *Entity2* に置き換える。

[1]太字が述語を、ARGn（n = 1, 2, ...）が述語の項を、MODARG が修飾先の述語項構造を示す。四角で囲まれた数字は共有構造を示す。

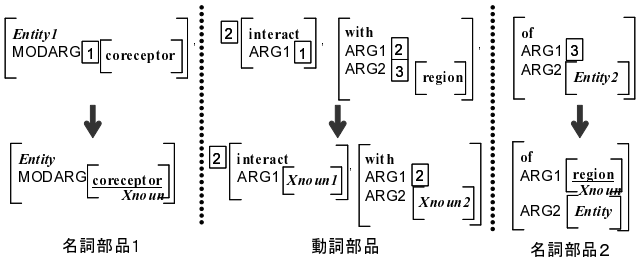



図 4: 述語項構造パターンの分割例

### 2.3 述語項構造パターンの分割

パターンをさらに一般化するため、動詞を 1 つだけ含むようなパターンを動詞（＋前置詞）部分（**動詞部品**）および名詞部分（**名詞部品**）に分割する。現在の実装では動詞を 2 つ以上含むあるいは動詞を全く含まないパターンは分割の対象外としている。

図 4 に例を示す。まず “interact with”の部分を動詞部品とし、その前後をそれぞれ名詞部品とする。動詞部品中の “coreceptor”（1）と “region”（3）は名詞部品への連結部であり、変数 *Xnoun* に変換する。また左の名詞部品中の “coreceptor” に、*Xnoun* と結合できるマークを付ける。右の名詞部品中の “region”にも同様にする。

このようにして得られた動詞部品と名詞部品のすべての組み合わせを、パターンとする。

### 2.4 述語項構造パターンの選別

このようにして抜き出されたパターンの中には実際には不適切なものも含まれる（実験と分析の項参照）ため、それらを除くようパターンの選別を次のようにして行う。ここまでに獲得したパターンを元の学習コーパスに適用し、その結果中の True Positive（TP）と False Positive（FP）の数を数える。TP − FP が一定値 $\theta$ 以下であるようなパターンは、不適切として削除する。

## 3 実験と分析

実験では、述語項構造パターンの学習コーパスおよびテストコーパス双方として Aimed[4] を用いた。用いたのはタンパク質間相互作用およびタンパク質名についてタグ付けされた 199 件の MEDLINE アブストラクトである。既存研究 [4] では 230 件のアブストラクトを用い、表層単語列のパターンを機械学習で構築する手法で適合率約 55%、再現率約 40%（ただし後述する Abstract Unit において）を得ている。今回の実験では、時間的コストの問題から長すぎる文あるいは複雑すぎる文（83 文）を除いた 1721 文を用いた。またタンパク質の named entity はすでに付与されているとした。

パターン適用による情報抽出の評価の基準には、Word Unit と Abstract Unit の 2 つを用いた。Word Unit では、相互作用するとタグ付けされているすべてのタンパク質の語が抜き出されなければならない。Abstract Unit では、相互作用するタンパク質の組はアブストラクト中のいずれかの文から抜き出されればよいとする。図 5 に、ベースライン（抽出のみの場合）、分割を行った場合、さらに選別を行った場合のそれぞれの 10-fold cross validation による結果を示す。選別を行った場合については $\theta$ を変えた結果をプロットしてある。図中には Word/Abstract Unit それぞれで F 値最大の点における適合率、再現率、F 値を示している。

図 5 に示された通り、まず述語項構造パターンの分割によって適合率をほぼ保ったまま再現率を上げている。再現率が上がったことから、パターンの分割によって学習コーパスに出現しなかったような動詞部品と名詞部品の組み合わせに対応できるようになったことが分かる。ここで適合率がほとんど下がらなかったことから、分割によって生じうる動詞部品と名詞部品の不適切な組み合わせはテストコーパスで出現することがあまりなかったと分かる。次にパターンの選別によって、再現率はやや下がるものの適合率を大きく上げている。先の分割導入時に適合率が下がらなかったことから、選別によって削除された不適切なパターンは分割の対象とならなかったものに多かったと考えられる。

表 1 に、実際に選別によって削除されたパターンの例を TP − FP の値が小さかった方から示す。例えば (1) は “Interactions of the *Entity1* protein with itself” という語句から抜き出されてしまった不適切なパターンである。この語句から適切なパターンを抜き出すためには、“protein”は単独でパターンとなるには不適切であることなどを知識としてシステムに組み込む必要がある。また (4)、(5) は動詞部品と名詞部品の不適切な組み合わせであるのに対して、(6) は構文解析が間違っていたために抽出された不適切なパターンであった。

## 4 まとめ

本稿では、情報抽出のための抽出ルールを、構文解析の結果である述語項構造のパターンとして構

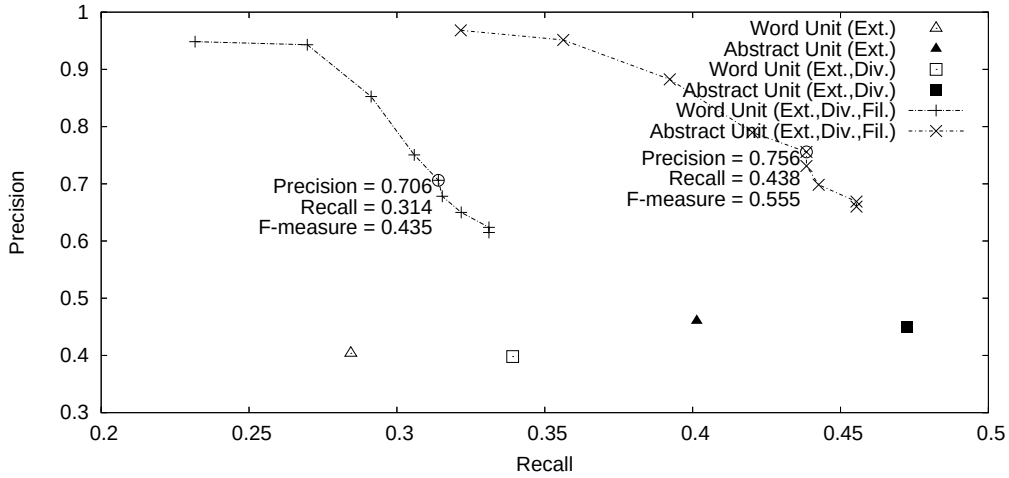



図 5: 抽出結果精度

| 分割されなかったパターン | | 分割されたパターン | | | |
|---|---|---|---|---|---|
| | | | 動詞部品 | 名詞部品 | |
| (1) | *Entity1* protein | (4) | *Xnoun1* be *Xnoun2* | $\underline{Entity}_{Xnoun}$ | complex$_{Xnoun}$ of *Entity* |
| (2) | *Entity1* complex | (5) | *Xnoun1* contain *Xnoun2* | $\underline{Entity}_{Xnoun}$ | *Entity* $\underline{\text{complex}}_{Xnoun}$ |
| (3) | domain of *Entity1* | (6) | *Xnoun1* induce *Xnoun2* | $\underline{Entity}_{Xnoun}$ | $\underline{Entity}_{Xnoun}$ |

表 1: 選別後削除された述語項構造パターン例

築する手法を提案した。述語項構造を用いることで表層上の単語列を用いるよりも一般化することができ、さらにパターンを分割することでより再現率を上げた。タンパク質間相互作用の抽出実験を行った結果、人手によるルール修正を行うことなく適合率 75.6%、再現率 43.8%を得た。今後残された課題としては、より高い再現率を得るために統語レベル以外の表現の変形（例えば “binding of *Entity1* to *Entity2*” と “binding of *Entity1* and *Entity2*” など）に対応すること、また自動獲得した抽出ルールの人手による改良があげられる。

## 参考文献


[1] Asako Koike, Yoshiyuki Kobayashi, and Toshihisa Takagi. Kinase Pathway Database: An Integrated Protein-Kinase and NLP-Based Protein-Interaction Resource. *Genome Research*, 13:1231–1243, 2003.

[2] Nikolai Daraselia, Sergei Egorov, Andrey Yazhuk, Svetlana Novichkova, Anton Yuryev, and Ilya Mazo. Extracting Protein Function Information from MEDLINE Using a Full-Sentence Parser. In *Proc. the Second European Workshop on Data Mining and Text Mining for Bioinformatics*, pages 15–21, 2004.

[3] Minlie Huang, Xiaoyan Zhu, Yu Hao, Donald G. Payan, Kunbin Qu, and Ming Li. Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*, 20(18):3604–3612, 2004.

[4] Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun Kumar Ramani, and Yuk Wah Wong. Comparative Experiments on Learning Information Extractors for Proteins and their Interactions. *Journal Artificial Intelligence in Medicine (Special Issue on Summarization and Information Extraction from Medical Documents)*, 2004.

[5] Ivan A. Sag and Thomas Wasow. *Syntactic Theory*. CSLI publications, 1999.

[6] Tsujii laboratory. Enju - A practical HPSG parser, 2005. http://www-tsujii.is.s.u-tokyo.ac.jp/enju/.

[7] Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The Penn Treebank: Annotating predicate argument structure. In *Proc. AAI '94*, 1994.