自然言語処理技術とオントロジーを利用した生物医学論文における多様な 2 項関係の抽出

秋谷 兼充[1]　牧野 貴樹[2]　クレイネス スティーブン[2]　高木 利久[1,2,3]
[1]東京大学大学院　新領域創成科学研究科　情報生命科学専攻　[2]東京大学　総括プロジェクト機構
[3]情報・システム研究機構　ライフサイエンス統合データベースセンター

１　はじめに

近年、自然言語処理技術を用いた、生物医学文献からの関係抽出の研究が盛んに行われている。たとえば、蛋白質や遺伝子とそれらの機能[1]や病気と疾患遺伝子の関係[2]など特定の 2 項関係を抽出するシステムの開発が行われてきた。しかし、このような関係抽出システムでは、実用に適した精度が得られているものの、2 項関係の種類が限定されているため、文献に含まれる情報全体を十分に把握できず、ユーザーの多様な要望に応えることができない。

一方、構文解析からテンプレートを作成し、2 項関係の関係抽出を行うシステム[3]では様々な種類の関係を抽出することが容易であるものの、テンプレートのみを抽出に用いているため、教師データの数が少ない場合、テンプレート数が少なくなり、高い精度が期待できない。精度を上げるために、テンプレート以外の手掛かりを利用して抽出する手法が必要とされていた。

そこで、我々は構文解析により得られたテンプレートに加え、テンプレート周囲の単語情報や関係のある用語に対応する意味のクラス情報を用いて機械学習することで、構文木から型付 2 項関係を抽出する手法を提案する。本手法で得られたテンプレートと学習された判別ルールを使うことで、新しい論文アブストラクトから型付 2 項関係を良い精度で自動抽出することができる。構文解析を行うことで、統語上の変形を吸収し、単語列よりも一般化されたテンプレートを得ることができる。構文部分木変換により再現率を高めることができる。また、機械学習を用いることでテンプレート以外の関係の特徴となる情報も利用でき、適合率を高めることができる。

２　手法

我々のシステムでは専門家によりエンティティとクラスがタグ付けされた論文アブストラクトから型付 2 項関係の抽出を行う。論文アブストラクト中に現れる生物医学用語はエンティティとしてタグ付けされ、エンティティが表す概念にオントロジー中で対応するクラスが付与されている。型付 2 項関係 $E_1 \overset{R}{\rightarrow} E_2$ は、source エンティティ E_1 と destination エンティティ E_2 とその２つのエンティティ間の関係の型 R からなるものである。例えば“TPL2 kinase regulates NF-κB.”という文があり、[TPL2 kinase]エンティティに enzyme クラスが、[NF-κB]エンティティに transcription factor クラスが対応する型付 2 項関係は [TPL2 kinase $\overset{\text{interacts\_with}}{\rightarrow}$ NF-κB]のようになる。

我々のシステムの概要を図１に示す。グレーの部分が本研究で開発したシステムである。このシステムでは、エンティティ名およびそれに対応するクラスがタグ付けされた論文アブストラクトと、そのアブストラクトに対して、専門家が付与した型付 2 項関係を教師データとして、テンプレートと学習された判別ルールを得る。これらを用いて、エンティティ名とそれに対応するクラスがタグ付けられた新しい論文アブストラクトが与えられると、自動的に型付 2 項関係を抽出する。システムの前処理として、構文解析を行い、統語上の変形を吸収した構文部分木を作成する。

2.1 前処理

前処理は訓練時およびテスト時で共通である。

2.1.1 構文解析

入力された論文アブストラクトは、構文解析器 Enju[4]を用いて構文解析を行い、構文木を得る。

2.1.2 構文部分木抽出

アブストラクトの文章中にタグ付けされているエンティティのペア $\langle E_1, E_2 \rangle$ に対して、構文木の主辞を結ぶ構文木中の経路部分を抜き出し、含まれる単語をすべて原形に置換して構文部分木 (Tree Fragment) 集合 $TF(E_1, E_2)$に格納する。エンティティ E_1, E_2を結ぶ経路が複数存在する場合には、各経路に対応する構文木を全て抜き出す。実際に、論文アブストラクトから構文部分木を作成した例を図２に示す。

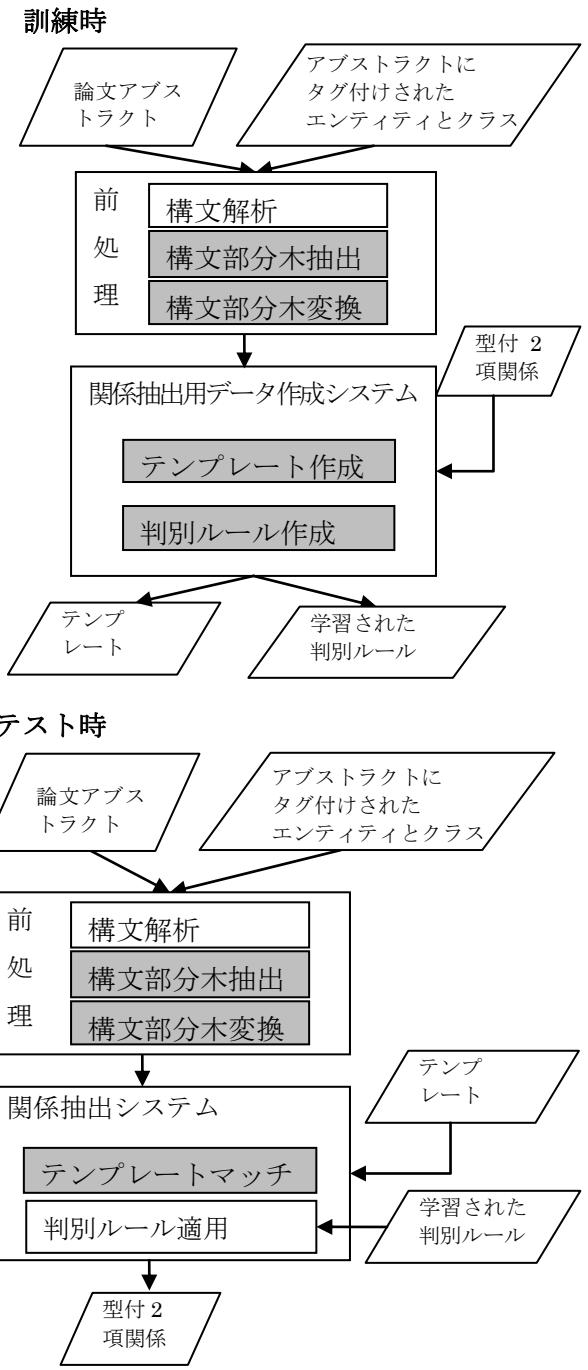

図 1． システム概要

2.1.3 構文部分木変換

次に、マッチングの再現率を上げるため、統語変換および動詞抽出変換を行う。これらのルールで追加された構文部分木は、元の構文木から抽出された構文部分木と同様に、その後の学習・テンプレートマッチの処理に利用される。これらのルールにより、適合率は低下するが、再現率が向上することが期待できる。

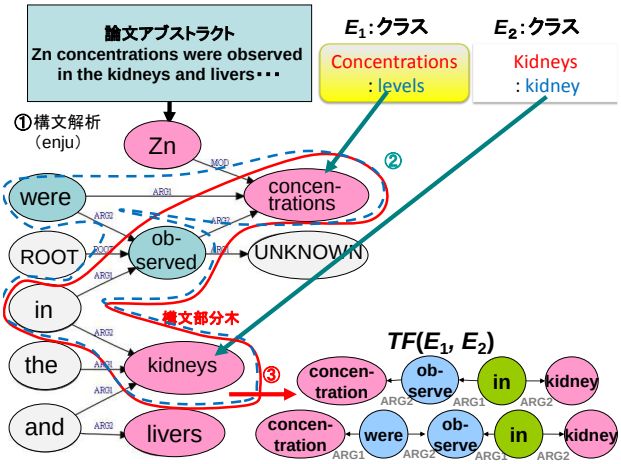

図 2． 構文部分木の作成

① 統語変換ルール：統語上の違いを吸収することで、テンプレートがさまざまな表現にマッチするようになり再現率の向上が期待できる。

(ア) of に対する統語変換：2 つの名詞 Zn concentrations からなる構文部分木が $TF(E_1,E_2)$に含まれる場合 concentrations of Zn という形の構文部分木を $TF(E_1,E_2)$に追加する。

(イ) 等位接続語に対する統語変換：concentrations were observed in kidneys and livers のような and を含む構文部分木が $TF(E_1,E_2)$に含まれる場合、concentrations were observed in livers という文に対応する構文部分木を $TF(E_1,E_2)$に追加する。

(ウ) be 動詞に対する統語変換：STIM1 is the molecule that regulates reaction のような、2 つの名詞にはさまれた be 動詞と他の部分木からなる構文部分木が $TF(E_1,E_2)$に含まれる場合、STIM1 regulates reaction のような文に対応する構文部分木を $TF(E_1,E_2)$に追加する。

② 動詞抽出変換：両端ノードと 1 つの動詞と他のノードを含む構文部分木の場合、他のノードを削除した構文部分木を $TF(E_1,E_2)$に追加する。この変換により 2 項関係の型 R を表現する動詞に注目し、動詞が一致する場合にはテンプレートがマッチするという効果が得られる。例を図 3 に示す。

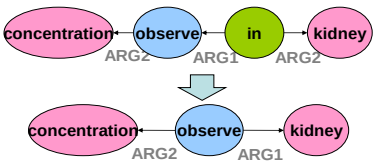

図 3　動詞抽出変換の例

2.2 関係抽出用データ作成システム

関係抽出用データ作成システムでは $TF(E_1,E_2)$ から各々の2項関係の型 R ごとのテンプレートと学習された判別ルールを得る。

2.2.1 テンプレート作成

各々の 2 項関係の型 R に対して、$E_1 \xrightarrow{R} E_2$ という関係が存在するすべてのエンティティペア $\langle E_1,E_2\rangle$ に関連付けられた $TF(E_1,E_2)$ に含まれる構文部分木すべてを、変換ルールを用いて構文部分木からテンプレートへ変換し、テンプレート集合 $T(R)$ に格納する。テンプレートは構文部分木のノード情報（単語の原形、品詞またはエンティティに対応するクラス）とエッジ情報（単語同士の関係）からなり、テストデータを構文解析した構文部分木とマッチングすることで、テストデータ中の型付2項関係の候補を抽出するために使われる。具体的には以下のようなルールで、構文部分木に含まれるノード情報を適切に単語の原形、品詞またはクラスに置換する。

① 構文部分木において両端のエンティティノード以外にコンテンツワード（動詞・名詞・形容詞）をノードに含む場合、両端ノードを品詞に置換し、中間ノードに含まれるコンテンツワードを型の抽出に使う。例を図4(a)に示す。

② 構文部分木において両端のエンティティノード以外にコンテンツワードをノードに含まない場合、以下の 4 パターンのテンプレートを生成する。1) E_1 を品詞に置換し、E_2 は単語の原形のまま。2) E_2 を品詞に置換し、E_1 は単語の原形のまま。3) E_1 を品詞に、E_2 を対応付けられたオントロジーのクラスに置換。4) E_2 を品詞に、E_1 を対応付けられたオントロジーのクラスに置換。①と同様に両端を品詞に置換すると、マッチする対象が膨大になり、精度が大幅に落ちるため、片方のノードで原形またはクラス情報を保持することで、マッチする対象を絞り込む。例を図4(b)に示す。

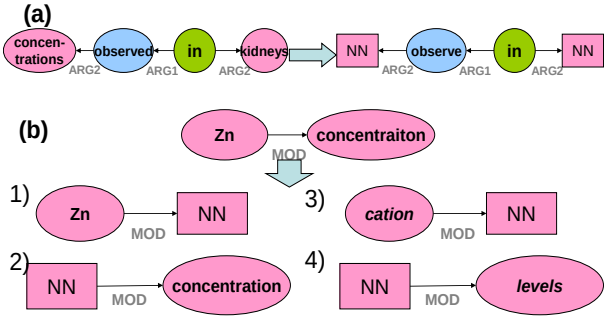

図4．テンプレート作成例

2.2.2 判別ルールの作成

訓練時に各々の型ごとに訓練データから型付 2 項関係の特徴を表す情報を取り出し、学習された判別ルールを作成する。ルールを作成するのに用いられる機械学習用データには訓練データから獲得したある2項関係の型 R に対するテンプレート集合 $T(R)$ をもう一度同じ訓練データに適用し、その結果、マッチした構文部分木と、その部分木に対応するエンティティ、クラスを用いる。テンプレートマッチが成功した構文部分木に対応するエンティティペア $\langle E_1,E_2\rangle$ の中で、訓練データ中に $E_1 \xrightarrow{R} E_2$ の関係が実際に存在する場合を正例、それ以外を負例として機械学習を行う。用いた特徴は表1に示すように、以下の3つのグループに分けられる。

- テンプレート情報：テンプレート集合 $T(R)$ に含まれる各々のテンプレートがどの程度の精度で型 R の関係を抽出できるかを学習することで、効果のあるテンプレートを識別し、関係抽出の精度を向上させる。
- クラス情報：オントロジーに含まれるクラスを学習特徴に用いる。クラス同士の関係がどの程度の精度で型 R を抽出できるかを学習することで、効果のあるクラスのペアを識別し、関係抽出の精度を向上させる。
- 周辺ノード情報：構文部分木の周辺の単語情報がどの程度の精度で型 R を抽出できるか学習するため、判別器用データに含まれる構文部分木の周辺ノードの単語ごとに True-Positive（TP）と False-Positive（FP）の数を数え上げる。TP/(TP+FP)または FP/(TP+FP)が 0.8 以上かつ TP+FP 数が多いものを source エンティティ、destination エンティティ、中間ごとに5つあげ、周辺ノード情報とする。

表1．機械学習に用いた特徴

Group of Features	Features	Values	Number of Features
テンプレート情報(tmp)	エンティティペア $\langle E_1,E_2\rangle$ を結ぶ構文部分木のいずれかにマッチするテンプレート	Binary	$T(R)$に含まれるテンプレートの個数
クラス情報(cl)	判別器用データに含まれる source エンティティに対してタグ付けされたクラス	Binary	オントロジーが持つクラスの個数
	判別器用データに含まれる destination エンティティに対してタグ付けされたクラス	Binary	オントロジーが持つクラスの個数
周辺ノード情報(wd)	source エンティティに隣接する構文部分木に含まれないノード情報	Binary	5
	destination エンティティに隣接する構文部分木に含まれないノード情報	Binary	5
	source エンティティと destination エンティティにはさまれた中間の構文部分木に隣接する構文部分木を含まないノード情報	Binary	5

2.3 関係抽出システム

全ての2項関係の型Rに関連付けられたテンプレート集合 $T(R)$とテストデータから抽出した構文部分木集合 $TF(E_1,E_2)$との間で、テンプレートマッチを行い、エンティティペア〈E_1,E_2〉がマッチしたら型付 2 項関係 $E_1 \xrightarrow{R} E_2$ を正解候補に格納する。その後、学習された判別ルールにより正例に分類されたものだけを抽出し、システムの出力とする。

3　実験と結果

実験では、EKOSS[5]で使われているUnivOnText オントロジーに基づいて学習を行った。UnivOnText オントロジーには 734 個のクラスと 115 個の型からなっている。前処理で逆の関係（たとえば、has_location と location_of）を用いて、型の統一を行い、67 個の型に絞った。実験に用いたコーパスは381個の論文アブストラクトと、それに対して、クラスと型を用いて専門家がタグ付けした 4770 個の型付 2 項関係である。既存研究[3]では訓練用テンプレート集合を訓練データに適用し、その結果中の TP と FP の数を数え、TP-FP が一定値 θ 以下であるようなテンプレートは不適切として削除し、残りの選別したテンプレートを用いてマッチングを行っている。今回の実験では、(A) テンプレートマッチのみで関係抽出を行った場合と（B）機械学習を使わない既存研究の手法、(C) 全てのテンプレートを用いてマッチングを行い、学習された判別ルールを用いて関係抽出を行う手法を比較した。さらに、機械学習においてどのような情報が関係抽出に効果があるか調べるために、3つの特徴グループの組み合わせ、テンプレートのみ（tmp）とテンプレートとクラス（tmp+cl）とテンプレートとクラスと周辺ノード（tmp+cl+wd）で関係抽出を行い、比較した。機械学習にはデータマイニングツールである Weka[6]に含まれる SVM[7]を用いた。

表2に 10 分割交差検定による結果を示す。既存研究の手法を行った場合は θ を変えて F-measure が最も良かった結果を載せた。

表２　関係抽出実験結果

	適合率(precision)	再現率(recall)	F-measure
(A)	0.081	0.277	0.125
(B)	0.193	0.179	0.186
(C)tmp	0.447	0.081	0.137
(C)tmp+cl	0.487	0.127	0.201
(C)tmp+cl+wd	0.398	0.073	0.124

表２に示された通り、本手法の最も良い結果と既存研究手法の最も良い結果を比較すると、再現率はやや下がるものの、適合率が高い結果になり、F-measure が向上した。このことから、提案するテンプレートマッチと機械学習の組み合わせによる手法の有効性が確認できた。また、機械学習においてクラス情報が関係抽出を行う効果が大きいことが分かる。機械学習に周辺ノードを用いた場合、用いない場合と比較して適合率および再現率が下がっているのは、訓練データが少ないため、型を識別する上で特徴となる単語を抽出できなかったためと考えられる。

4　まとめ

本研究では、関係抽出のために、構文解析結果から統語上の変形を吸収したテンプレートとテンプレートやクラスといった情報を学習した判別ルールを用いる手法を提案した。テンプレートマッチの後、機械学習を用いることで、適合率を上げられることを示した。少ない訓練データから多様な関係抽出実験を行った結果、適合率 48.7%、再現率 12.7%を得た。今後の課題として、適合率を上げるために、クラスの階層構造を利用することや自動抽出した周辺ノード情報を人手により選別して、収集した単語を判別ルールに用いることがあげられる。また、再現率を上げるためには、テンプレートや判別ルールの元になる訓練データの増加が必要不可欠であると考えられる。

参考文献

[1] Koike A et al. Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics*, **21**:1227-1236, 2006.

[2] Chun H et al. Extraction of Gene-Disease Relations from MEDLINE Using Domain Dictionaries and Machine Learning. Pacific Symposium on Biocomputing. **11**: 4-1, 2006.

[3] Yakushiji A, Miyao Y, Ohta T, Tateisi Y, Tsujii J. Automatic Construction of Predicate-argument Structure Patterns for Biomedical Information Extraction. In *Proc. 2006 Conference on Empirical Methods in Natural Language Processing,* pp284-292, Sydney, Australia, 2006.

[4] Torisawa K and Tsujii J.　An HPSG-Parser for Automatic Knowledge Acquisition. In *Proc. 4th International Workshop on Parsing Technologies (IWPT)*, pp. 250-251, Prague, Czech republic, 1995 .

[5] Kraines S et al. EKOSS: A knowledge-user centered approach to knowledge sharing, discovery, and integration on the Semantic Web. *Journal of Information Processing and Management*, **50**:322-342, 2007.

[6] Witten I, Franck E, Trigg L, Hall M, Holmes G and Cunningham S.　Weka: Practical machine learning tools and techniques with Java implementations. In *Proc. ANNES'99 International Workshop on emerging Engineering and Connectionist-based Information Systems*, pp.192–196, 1999.

[7] Bernhard E, Isabelle M, Vladimir N, A Training Algorithm for Optimal Margin Classifiers. 5th COLT, pp.144-152, 1992.