お詫びと訂正

本論文の言語処理学会15回年次大会論文集に収録した版において、実験における比較対象が引用した研究内容と無関係の別の手法になっておりました。謹んでお詫びさせていただくとともに、訂正した原稿を配布させていただきます。関係者の皆様には大変ご迷惑をおかけして申し訳ございませんでした。言語処理学会事務局様には、訂正原稿の配布に関しまして大変お世話になりました。心より感謝申し上げます。

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二項関係の抽出

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、論文数の爆発的な増加を背景として、論文からの知識自動抽出の必要性が高まっている。生物医学分野における知識の多くはタンパク質間相互作用[1]に見られるように二項関係として表すことができる。しかし、二項関係には様々な種類があるため、その多様性を統一的に扱うためには、二項関係に型を付与し、型付二項関係として知識を表現することが有効である。

生物医学分野における二項関係抽出は、タンパク質間相互作用や病気と疾患遺伝子の関係[2]など特定の物質や事象において一定の成果を上げている。しかし、このような特定の事象だけではない、多様な型付二項関係を自動抽出するためには、文章中に現れる語句同士を結び付けている様々な表現と、その表現が意味する二項関係との複雑な対応関係が必要となる。このために、自動で抽出パターンを作成し、抽出パターンと機械学習を組み合わせて様々な二項関係を抽出できる手法[3]を適用することが考えられるが、高い精度を得ることは難しかった。この理由のひとつとして、型付二項関係抽出の精度を上げるために必要なパターンマッチした文の文脈を考慮していないことが挙げられる。

そこで、我々は様々な型に対応した精度の高い型付二項関係の抽出を目的とし、論文アブストラクトを構文解析し、得られたパターン（述語項構造の述語を結合した抽出ルール）を利用したパターンマッチに加え、コーパス[4]に含まれる生物医学用語に対応するクラスのオントロジー（クラス階層）情報を機械学習に用いた関係抽出システムを提案する。本研究の特徴は機械学習にパターン情報に加えオントロジー情報を用いることで型付二項関係抽出を行うシステムの性能を上げることができる点である。

2　手法

本システムでは専門家によりエンティティとクラスが付与された論文アブストラクトから型付二項関係の抽出を行う。具体的には、入力の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マークされ、各々のエンティティには、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。出力となる型付二項関係 $E_1 \xrightarrow{R} E_2$ 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 クラスが対応する場合、この文の意味を表す型付二項関係は [TPL2 kinase $\xrightarrow{\text{interacts\_with}}$ NF-κB]のようになる。

本システムの概要を図1に示す。本システムでは、エンティティ名およびそれに対応するクラスが付与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二項関係を訓練データとして、パターンと学習された判別器を得る。これらを用いて、エンティティとそれに対応するクラスが付与された新しい論文アブストラクトが与えられると、自動で型付二項関係を抽出する。

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。構文解析をし、統語上の変形を吸収した raw パターンを得る。

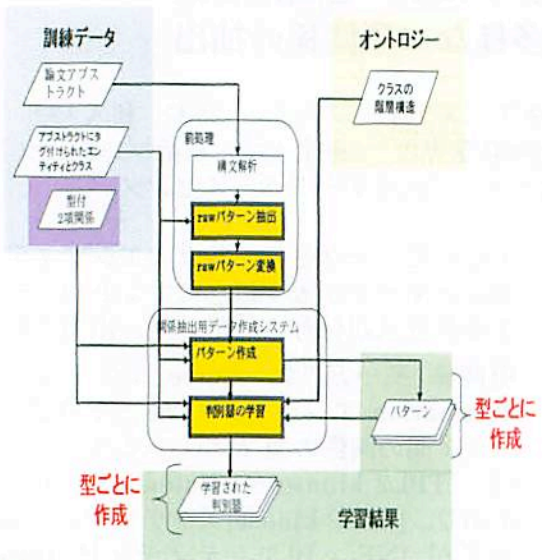
2.1.1　構文解析

入力された論文アブストラクトは、構文解析器 Enju[5]を用いて構文解析を行い、述語項構造の集合を得る。

2.1.2　raw パターン抽出

エンティティペア $\langle E_1, E_2 \rangle$ の E_1、E_2 を両端として結ぶ述語項構造集合中の経路部分を抜き出し、含まれる単語を全て原形に置き換えて raw パターン集合 $RP(E_1, E_2)$ 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 raw パターンを全て抜き出す。実際に、論文アブストラクトから raw パターンを作成した例を図2に示す。

訓練時



テスト時

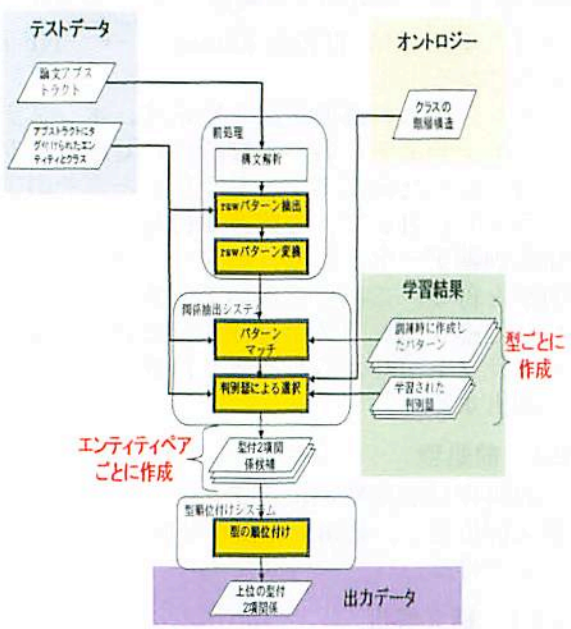


図１．システム概要

2.1.3 raw パターン変換

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 raw パターンは、元の述語項構造集合から抽出された raw パターンと同様に、その後の学習・パターン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。

① 統語変換ルール：統語上の違いを吸収することで、パターン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。

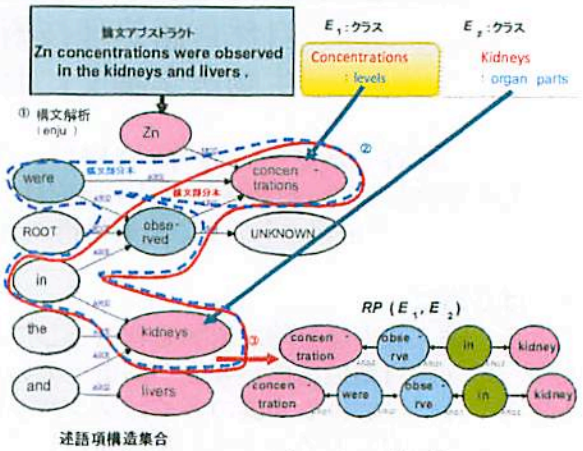


図２．raw パターンの作成

(ア) of に対する統語変換：２つの名詞 Zn concentrations からなる raw パターンが $RP(E_1,E_2)$に含まれる場合 concentrations of Zn という形の raw パターンを $RP(E_1,E_2)$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 raw パターンが $RP(E_1,E_2)$ に含まれる場合、concentrations were observed in livers という文に対応する raw パターンを $RP(E_1,E_2)$ 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 raw パターンが $RP(E_1,E_2)$に含まれる場合、STIM1 regulates reaction のような文に対応する raw パターンを $RP(E_1,E_2)$ 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 raw パターンの場合、他のノードを削除した raw パターンを $RP(E_1,E_2)$に追加する。削除した側のエッジを任意のエッジにマッチするものに置き換える。この変換により二項関係の型 R を表現する動詞に注目し、動詞が一致する場合にはパターンがマッチするという効果が得られる。例を図３に示す。

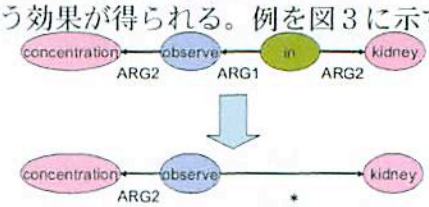


図３　動詞抽出変換の例

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 $RP(E_1,E_2)$ から各々の二項関係の*型* R ごとのパターンと学習された判別器を得る。

2.2.1 パターン作成

各々の二項関係の*型* R に対して、$E_1 \overset{R}{\rightarrow} E_2$ 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 $RP(E_1,E_2)$ に含まれる raw パターンすべてを、変換ルールを用いて raw パターンからパターンへ変換し、パターン集合 $T(R)$ に格納する。パターンは raw パターンのノード情報（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 raw パターンとマッチングすることで、テストデータ中の型付二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、raw パターン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 raw パターンでは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 raw パターンでは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 を対応付けられたクラスに置換。4) E_2 を品詞に、E_1 を対応付けられたクラスに置換。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る。例を図 4 (b)に示す。

2.2.2 判別器の学習

訓練時に各々の型ごとに訓練データから型付二項関係の特徴を表す情報を取り出し、判別器を学習する。機械学習用データには訓練データから得たある二項関係の*型* R に対するパターン集合 $T(R)$ 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 raw パターンと、その raw パターンに対応するエンティティ、クラスを用いる。パターンマッチが成功した raw パターンに対応するエンティティペア $\langle E_1,E_2\rangle$ の中で、実際に $E_1 \overset{R}{\rightarrow} E_2$ の関係が訓練データ中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

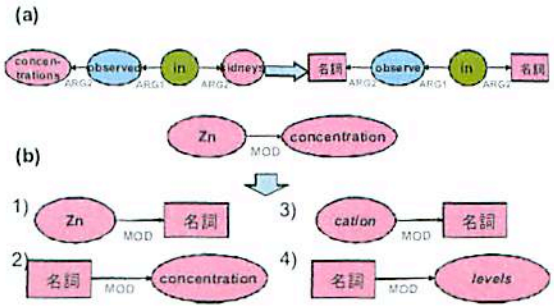


図 4．パターン作成例

以下の 4 グループあり、表 1 にまとめる。

- パターン情報(pt)：パターン集合 $T(R)$ に含まれる各々のパターンがどの程度の精度で*型* R の関係を抽出できるかを学習して、効果のあるパターンを識別する。
- クラス情報(cl)：エンティティのクラスを素性に用いる。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して、効果のあるクラスのペアを識別する。
- オントロジー（クラス階層）情報(ot)：オントロジーに含まれるクラスの階層構造を学習に用いる。つまり、上記のクラス情報に加え、クラスの上位クラス全てを素性に用いる。これにより、単にエンティティペアに含まれるクラス同士の関係のみならず、上位クラス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ができる。
- 周辺単語情報(wd)：raw パターン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、機械学習用データに含まれる raw パターンの周辺単語ごとに、あるエンティティペアがあったときにその周辺単語が存在する確率とエンティティペアが型 R になる確率の相互情報量を求める。source エンティティ、destination エンティティ、中間ごとに相互情報量が一番高い単語を素性に用いる。

2.3 関係抽出システム

各々の二項関係の型 R に関連付けられたパターン集合 $T(R)$ とテストデータから抽出した raw パターン集合 $RP(E_1,E_2)$ との間でパターンマッチを行い、エンティティペア $\langle E_1,E_2\rangle$ がマッチしたら型付二項関係 $E_1 \overset{R}{\rightarrow} E_2$ を正解候補とする。その後、エンティティペア $\langle E_1,E_2\rangle$ と学習された判別器により出力された予測値と、型を正解候補の型付二項関係集合 $C(E_1,E_2)$ に格納する。

2.4 型順位付けシステム

正解候補の型付二項関係集合 $C(E_1,E_2)$ に格納されている各々のエンティティペア $\langle E_1,E_2\rangle$ に対応する型の予測値を比較して、予測値の高い順番に型

の順位付けを行う。上位の型をエンティティペア〈E_1,E_2〉の型として、型付二項関係を出力する。

表１．機械学習に用いた素性

Features	Contents	Values	Number of Features
パターン情報(pt)	エンティティペアを結ぶrawパターンのいずれかにマッチするパターン	Binary	$T(R)$に含まれるパターンの個数
クラス情報(cl)	機械学習用データに含まれるsourceおよびdestinationエンティティのクラス	Binary	オントロジーが持つクラスの個数
オントロジー情報(ot)	機械学習用データに含まれるsourceおよびdestinationエンティティのクラスとそれらの上位クラス全部	Binary	オントロジーが持つクラスの個数
周辺単語情報(wd)	source、destinationエンティティおよび2つのエンティティにはさまれた中間rawパターンに隣接する単語	Binary	3

3　実験と結果

実験では、EKOSS[4]で使われているUnivOnTextオントロジーとコーパスを利用して、関係抽出システムの評価を行った。具体的には、UnivOnTextオントロジーに含まれる1774個のクラスと65個の型を用いた。また、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3288個の型付二項関係を用いた。機械学習に用いる素性を、①パターン情報(pt)のみ、②クラス情報(cl)のみ、③オントロジー情報(ot)のみ、④周辺単語情報(wd)のみ、⑤pt+otの組み合わせ、⑥pt+ot+wdの組み合わせ、の6通りで実験した。既存研究[3]は、本研究のシステムよりも洗練されたパターン抽出手法を使っているものの、機械学習ではパターン情報のみを機械学習に用いているため、ptのみに近い傾向の結果が出ると考えられる。機械学習にはWeka[6]に含まれるSVM[7]を用いた。

図５に10分割交差検定による結果をPrecision-Recall（PR）カーブで表し、曲線下面積(AUC)を示す。実験結果から、パターン情報(pt)とオントロジー情報(ot)を組み合わせたときが、AUCが最も良かった。特に、再現率が0.1以上の領域で、ptのみでは適合率が急速に落ちるが、pt+otではあまり適合率が下がらない傾向が読み取れる。これは、構文構造を表すパターンと語句の意味を表すオントロジーのよさを組み合わせた結果であると考えられる。このことから、本研究で提案するパターン情報とオントロジー情報を機械学習に用いる手法は、型付二項関係を抽出する上で、効果があると考えられる。

単語情報を用いたときにAUCが下がったのは、SVMが出力した各々の型ごとの予測値を混ぜてPRカーブを求めたためだと考えられる。各々の型ごとのAUCを比較すると、単語情報を加えたほうが、AUCが上がっている型が多かった。

オントロジー情報とクラス情報を比較すると、オントロジー情報のほうが、AUCが高かった。これは、クラスのみではなくその上位クラス同士の関係を考えることが様々な関係抽出を行う上で、有効であると考えられる。

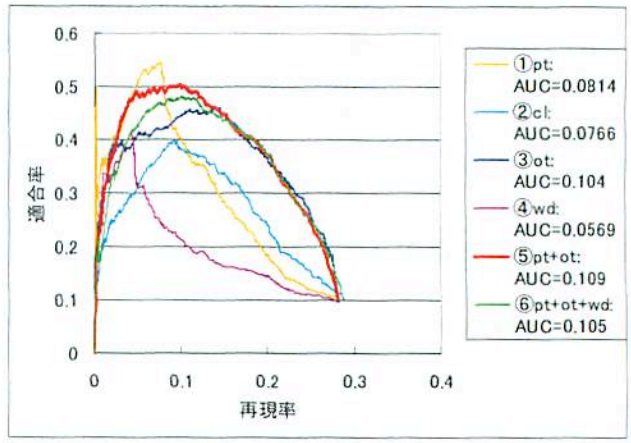


図５．関係抽出実験結果

4　まとめ

本研究では構文解析とオントロジー（クラス階層）情報を利用して機械学習を行う型付二項関係の自動抽出システムを提案した。実験の結果、本研究で提案するパターン情報とオントロジー情報を機械学習に用いる手法は、型付二項関係を抽出する上で有効であることが示された。

参考文献

[1] Fundel F, Kuffner R, Zimmer R, RelEx-Relation extraction using dependency parse trees. *Bioinformatics*, **23**:365-371 2007

[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. I*n Proc. 2006 Conference on Empirical Methods in Natural Language Processing*, pp284-292, Sydney, Australia, 2006.

[4] 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.

[5] Torisawa K and Tsujii J. Compiling HPSG-style grammar to object-oriented language. *In the Proceedings of NLPRS*, pp. 320-325, 1995 .

[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.