蛋白質相互作用抽出への転移学習の応用

三輪 誠　　辻井 潤一
東京大学　東京大学/マンチェスター大学/NaCTeM

概要

文献からの蛋白質相互作用の自動抽出は関係抽出のよい題材であるとともに，その作用は生物学的プロセスの解析に必須な情報である．この抽出の中で，文における 2 つの蛋白質間の関係抽出は 2 値の分類問題として広く研究され，様々なアノテーションポリシによるコーパスが公開されている．我々は複数の構文解析器・カーネルを元にした素性を利用したサポートベクタマシンによる蛋白質相互作用抽出システムを作成した．またそれぞれのアノテーションの違いを低減しつつ他のコーパスに利用する転移学習により，その抽出の精度の向上を目指した．結果として，既存のシステムと同等かそれ以上の精度で蛋白質相互作用を抽出できることが分かった．

1　はじめに

蛋白質相互作用 (PPI，Protein-Protein Interaction) は生物学的プロセスの解析に必須な情報である．現在多くの生物学データベースが人手で作成・更新されているが，生物学の文献は日々増え続けており，この作業は非常に手間のかかるものとなっている．この手間を軽減するために文献からの PPI の自動抽出が必要とされている．

このような情報抽出のうち 1 文中の 2 つの蛋白質ペアの相互作用抽出 (sentence-based pair-wise PPI extraction) は最も基本的な問題であるとともに関係抽出のよい題材でもあるため，そのペアが相互作用しているか否かの 2 値の分類問題として広く研究されている．このために共起関係を元にした単純なシステムから自然言語処理ツールを利用した機械学習システムまで数多くのシステムが提案されている．また，この問題を対象とした様々なアノテーションポリシによるコーパスが現在公開されている．これらのコーパスはそれぞれのコーパスについての学習曲線は増加しており，学習に十分な量ではない [1] ため，コーパスの同時利用が望まれる．しかし，それぞれ取り扱う PPI の定義が異なるため表 2 に示す結果のように容易に他のコーパスに流用することができない．

XPG$_{\mathrm{p1}}$ protein interacts with multiple subunits of **TFIIH**$_{\mathrm{prot}}$ and with **CSB**$_{\mathrm{p2}}$ protein.

図 1　相互作用している蛋白質ペア (p1，p2) を含んだ文の例 (AIMed PMID 8652557，9 文目 3 番目のペア)

この抽出問題のために，我々は複数の構文解析器 (パーザ) の情報を利用した複数のカーネルを元にした素性を用いた線形サポートベクタマシン (SVM) による PPI 抽出システムを作成した．また，それぞれのコーパスの違いを低減しつつ他のコーパスに利用し，学習データ不足の問題を軽減するために，転移学習 (transfer learning) を提案・適用し，その抽出の精度向上を目指した．結果として，既存のシステムと同等かそれ以上の精度で PPI を抽出できることが分かった．

2　PPI 抽出システム

我々のシステムは文中の蛋白質ペア (p1，p2) が相互作用するか否かを当てる問題を対象とする．文の例を図 1 に示した．我々はこの問題に対し，文を複数のパーザで解析し，素性に変換し，分類器を用いてモデルを作成する．

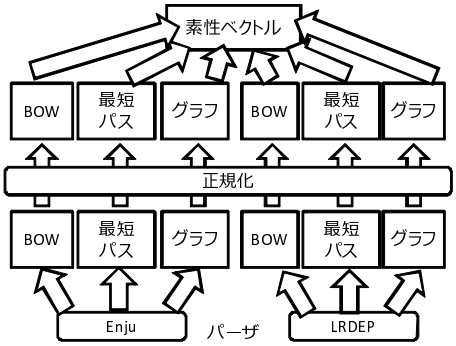

図 2　素性ベクトル

2.1　パーザ

我々は高速な係り受け解析を行う LRDEP と深い解析を行う Enju の 2 つのパーザ [4] を利用する．パーザから得られる構造としては LRDEP の係り受け構造と Enju の述語項構造を利用する．

2.2　素性

我々は利便性と高速化のため 3 種類のカーネル [4] とほぼ同等の線形の素性を作成する．具体的にはパーザごとに 3 種類の素性を抽出し，それぞれを個別に正規化したものを並べることによって図 2 のように素性ベクトルを作成する．一般性のために蛋白質の名前は用いず，p1 は ENTITY1，p2 は ENTITY2，その他の蛋白質は PROTEIN という単語として扱う．

Bag of words (BOW)　BOW 素性はパーザから得られる単語の原型 (lemma) に位置情報，頻度情報を付加したものである．位置情報としてはペアの前・間・後ろの 3 つを用いた．図 1 の BOW 素性を図 3 に示した．

最短パス　最短パス素性はペア間の構文構造における最短パス上に現れる関係の情報を表現したものである．この素性としては 2 つの単語とその間の関係を元に作成した v-walk と 1 つの単語とその周囲の関係 2 つを元に作成した e-walk，そしてそれらの部分木全てを列挙して用いた．図 1 の最短パスとそれに含まれる v-walk，e-walk を図 4 に示した．

PROT_M:1, and_M:1, interact_M:1, multiple_M:1, of_M:1, protein_M:1, subunit_M:1, with_M:2, protein_A:1

図 3　位置情報と頻度情報を含む BOW 素性 (B:Before，M:in the Middle of，A:After)

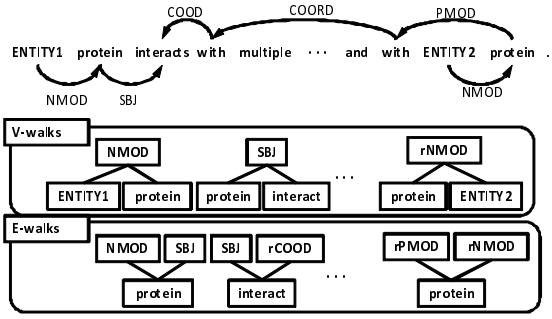

図 4　最短パスとそれに含まれる v-walk，e-walk

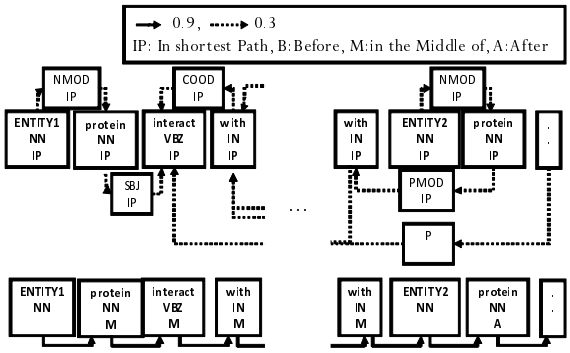

図 5　グラフ素性を抽出するための構文構造グラフとシーケンスグラフ

グラフ　グラフ素性は全経路グラフカーネルを線形の表現にしたものである．それぞれの素性は構文構造をグラフにした構文構造グラフまたは文における単語の並びをグラフにしたシーケンスグラフにおける任意の 2 ノード間の距離を表す．図 1 から得られる 2 つのグラフを図 5 に示した．ノードとは単語または関係である．経路の重みは経路上のエッジの重みを全て掛け合わせることで，距離はノード間の全ての経路についてその重みを足しあわせることで計算される．

2.3 分類器

分類器としては線形 L2 ソフトマージン SVM (L2-SVM) [3] を用いる．

3 転移学習

転移学習 (transfer learning) [5] とは他の問題の情報を利用し，対象問題の学習の精度向上を測る枠組みである．我々は比較的単純な 2 つの転移学習手法 Transfer SVM (TrSVM) ，ソフトマージン TrAdaBoost (SMTrAdaBoost) について提案・評価する．

TrSVM は L2-SVM のペナルティパラメータ C を対象問題と他の問題 (コーパス) で異なる値に設定することにより，問題間の差異を軽減しつつ対象問題の学習精度の向上を測る．

$$\min_{\boldsymbol{w}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C_s\sum_{i=1}^{ls}\xi_i + C_t\sum_{j=1}^{lt}\xi_j \qquad (1)$$

ここで ξ は損失関数，他の問題・対象問題それぞれについて ls・lt は例の数，C_s　0・C_t　0 はペナルティパラメータである．

SMTrAdaBoost は TrAdaBoost [2] においてソフトマージン AdaBoost (SMAdaBoost) [6] を利用するようにしたものである．TrAdaBoost は AdaBoost の反復において分類に失敗する他の問題の例の重みを小さくすることで，対象問題と異なる例の影響を小さくしようとする．SMTrAdaBoost における他の問題のパラメータはソフトマージン項がない場合に TrAdaBoost と同じとなるよう調整する．SMTrAdaBoost の弱学習器としては L2-SVM を用いる．

4 評価

評価には AIMed・BioInfer・HPRD50・IEPA・LLL (表中 A・B・H・I・L) の 5 つのコーパスを用い，アブストラクト単位での 10 分割交差検定 (10-fold cross valudation (CV)) で評価した．比較のため交差検定の分割方法は [1] と同じにした．SVM の学習には LIBLINEAR [3] の DCDL2 を用いた．言及がない限り実験ではペナルティパラメータは訓練データでの 10-fold CV で調整し，分類器の閾値は平均 F-score の最も高くなる値とした．これは若干楽観的な評価であるが，評価結果のよい予測といえる [4].

4.1 PPI 抽出システムの評価

表 1 に PPI 抽出システムの評価を示した．IEPA を除いた全てのコーパスにおいて，我々のシステム以外で最高の精度を出しているシステム [1] よりも高い精度で分類できている．

4.2 転移学習の評価

表 2 に示したのは，他のコーパスで学習したモデルを対象コーパスに適用した結果である．C は 1 とした．IEPA と LLL を除いた全てのコーパスペアについて，対象コーパスでのモデルの予測結果ほどの分類性能を得られていない．

表 3 に 3 章に示した TrSVM を行った結果を示した．これより多くのデータにおいて他のコーパスが対象コーパスの予測結果を向上させている．表 1 に表 3 において F-score が最高であったコーパスペ

表 2　異なるコーパスで学習したモデルの F-score (行が予測対象データ，列がモデル，モデルとデータが同じものは 10-fold CV の結果，C は 1 に固定)

	A	B	H	I	L
A	63.5	49.9	44.8	40.7	35.9
B	53.7	66.6	50.6	54.5	49.9
H	68.8	69.7	74.2	67.2	63.1
I	67.3	71.3	67.9	74.2	67.5
L	71.8	77.7	72.9	83.4	79.6

表 3　TrSVM の F-score (行が予測対象データ，列が他の追加したコーパス，モデルとデータが同じものは 10-fold CV の結果)

	A	B	H	I	L
A	64.2	64.0	64.7	65.2	63.7
B	67.9	67.6	67.9	67.9	67.7
H	71.3	71.2	69.7	74.1	70.8
I	74.4	75.6	73.7	74.4	74.4
L	83.2	85.9	82.0	86.7	80.5

表 1　PPI 抽出システムの性能評価 (TL は TrSVM について F-score が最高のもの．)

	正例	負例	F (SVM)	F (TL)	F [1]	AUC (SVM)	AUC (TL)	AUC [1]
AIMed	1000	4834	64.2	65.2	56.4	89.1	89.3	84.8
BioInfer	2534	7119	67.6	67.9	61.3	86.1	86.2	81.9
HPRD50	163	270	69.7	74.1	63.4	82.8	85.0	79.7
IEPA	335	482	74.4	75.6	75.1	85.6	87.1	85.1
LLL	164	166	80.5	86.7	76.8	86.0	90.8	83.4

アの結果を示した．結果より全てのコーパスにおいて，1 つのコーパスのみを使った結果よりも精度が高く分類できている．また SMTrAdaBoost を適用した結果を表 4 に示した．SMAdaBoost の精度は SVM よりも若干精度が低いものの，多くのデータについて表 3 の結果と同様，他のコーパスが予測結果を向上させている．

5　おわりに

本稿では蛋白質相互作用抽出システムと 2 つの転移学習を利用した複数の異なるコーパスを用いた学習について提案・評価を行った．

結果として現在提案されている蛋白質相互作用抽出システムと同等かそれ以上の精度で蛋白質相互作用の分類が可能なシステムを作成できた．またアノテーションポリシの異なる他のコーパスとの差異の影響を軽減することで，他のコーパスを用いて対象コーパスの予測精度を向上できた．

今後の課題としては，データを元にした素性空間の転移など例を直接用いた転移以外の手法，ラベルなしデータの利用がある．

表 4　SMTrAdaBoost の F-score (行が予測対象データ，列が他の追加したコーパス (all は対象データを除く全てのコーパス)，モデルとデータが同じものは 10-fold CV の結果)

	A	B	H	I	L	all
A	63.4	64.0	63.0	63.3	62.9	64.2
B	68.0	66.5	66.6	67.6	67.7	68.1
H	72.1	73.3	69.4	73.4	70.9	74.1
I	72.9	76.1	74.9	74.2	74.9	75.7
L	76.5	83.1	79.3	85.7	79.6	85.5

参考文献

[1] A. Airola, S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter, and T. Salakoski. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross corpus learning. *BMC Bioinformatics*, 2008.

[2] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Boosting for transfer learning. In *ICML*, pages 193–200, 2007.

[3] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

[4] M. Miwa, R. Sætre, Y. Miyao, T. Ohta, and J. Tsujii. Combining multiple layers of syntactic information for protein-protein interaction extraction. In *SMBM 2008*, pages 101–108, 2008.

[5] S. J. Pan and Q. Yang. A survey on transfer learning. Technical Report HKUST-CS08-08, Nov. 2008.

[6] G. Ratsch, T. Onoda, and K.-R. Muller. Soft margins for adaboost. *Machine Learning*, 42(3):287–320, 2001.