社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－MPS－72 (32)
2008－BIO－15 (32)
2008／12／18

関連データに基づく事例集合拡張による文献からの蛋白質名判定

宮西 一徳 † 尾崎 知伸 †† 大川 剛直 ‡

† 神戸大学大学院自然科学研究科 †† 神戸大学自然科学系先端融合研究環 ‡ 神戸大学大学院工学研究科

蛋白質構造解析に関する文献の増加に伴い，文献中から蛋白質名を自動的に判定する手法が求められている．しかし，蛋白質名を判定する際，訓練例が必ずしも十分ではなく，高い判定精度が得られない場合がある．また，利用可能な外部コーパス全体を訓練例に追加した場合，文の構造や語彙の違い等のため，負の影響が懸念される．そこで，蛋白質名判定の精度を向上させるため，訓練例に関連する文のみを外部コーパスより抽出し，訓練例集合を拡張する手法を提案する．実際に，提案手法を用いて，外部コーパスから文を抽出する実験を行ったところ，蛋白質名判定の精度の向上が確認できた．

Identification of Protein Names in Literature by Instance Set Extension

Kazunori Miyanishi † Tomonobu Ozaki †† Takenao Ohkawa ‡

† Graduate School of Science and Technology, Kobe University †† Organization of Advanced Science and Technology, Kobe University ‡ Graduate School of Engineering, Kobe University

As documents about protein structural analysis are increasing, a method of automatically identifying protein names in them is required. However the accuracy of identification is not high in the case of not enough training data set given. On the other hand, it may have a negative effect that a whole available corpus is added to training data set, because of differences of syntax structures or vocabulary. Then we propose a method to extend training data set by extracting related sentences from a corpus. In the experiment, it was confirmed that the accuracy was improved by the proposed method.

1 はじめに

蛋白質の機能情報は，創薬や生命現象の解明に有用であると考えられる．このような機能情報は，蛋白質構造解析に関する膨大な数の文献に記述されており，生物医学文献データベースである PubMed には 1800 万以上の文献データが登録されている．そこで，膨大な文献から機能情報を自動抽出することが期待されているが，その前処理として文中の蛋白質名を特定することが重要となる [1, 2]．

文献中の蛋白質名特定に関して様々な研究が行われている．これらの多くは単語の品詞や綴り等の特徴や，単語の前後関係を基にした機械学習によって，蛋白質名とそれ以外の単語を分類する手法に関するものである [3, 4, 5]．このような手法は，学習に対して十分な訓練例が与えられている条件下で蛋白質名判定を行う．そのため，訓練例が少ない場合には，高い判定精度が得られないという問題点がある．この問題点に対処するため，外部の関連するコーパスを利用することで，訓練例を補完し精度向上を図ることが考えられる．しかしながら，単純に外部コーパス全体を訓練例に加えると，コーパスと訓練例との間での文の構造や語彙の違い等のため，判定精度に悪影響を及ぼす恐れが生じる．そこで，本論文では，外部コーパスから関連する文のみを抽出し，訓練例集合を拡張することによって，蛋白質名判定の精度向上を図る手法を提案する．前述の通り，外部コーパスにおいて利用されている語彙は，必ずしも元の訓練例にお

けるものと同一でないため，蛋白質名そのものを手掛かりとして関連性のある文を抽出するのは困難である．そこで，蛋白質名の前後に頻出する語や，蛋白質名を含む構文構造をパターン化し，これをテンプレートとして利用することにより，蛋白質名判定に有用な文のみの抽出を実現する．

2 事例集合拡張の概要

2.1 事例集合とコーパス

本研究では，文中に含まれる蛋白質名や遺伝子名を表す単語を事例とする．さらに，文集合として，これらに対してタグ付けされたコーパスを利用する．このようなコーパスはそれぞれ特定の生物種を対象として構築されているものが多い．例えば，BioCreAtIvE 1 の Task 1B[6] で使用されるコーパスには “fly”, “mouse”, “yeast” の 3 種類がある．ここで，対象とする生物種によって，蛋白質名の語彙や略語を用いるなどの記述様式が異なる場合がある．そのため，異なる生物種を対象とするコーパス同士を単純に組み合わせると，蛋白質名判定に悪影響を及ぼすことが考えられる．“fly” と “mouse” に含まれる，略語表現を多用する場合とそうでない場合の文の例を以下に示す．太字の箇所が蛋白質名を表す．

“... the expression of these RNAs may reflect an early regulation of **Sxl** at the level of transcription ...” (PMID[1] : 2473007)

“Mouse **probasin** cDNA was isolated from a phage library, and the DNA sequence was determined.” (PMID : 10861744)

2.2 事例集合拡張

本研究では，訓練例が少ない場合に，他の利用可能な外部コーパスから事例を含む文を抽出し，訓練例を拡張して，蛋白質名の判定精度を向上させることを目的としている．事例集合拡張の概要を Fig. 1 に示す．単純に外部コーパス中の文を訓練例に追加すると，負の影響が懸念されるため，訓練例に関連する文のみを外部コーパスから抽出する必要がある．そこで，蛋白質名の周辺に現れやすい単語や構文情報を基に，訓練例に関連する文を外部コーパスから選択する．

[1] PubMed から参照される文献の識別番号

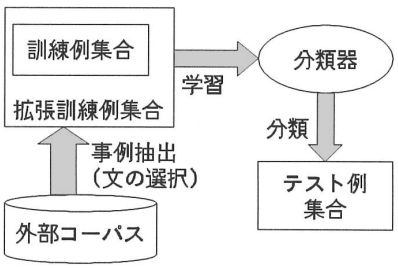

Fig. 1 事例集合拡張の概要

3 関連文の選択手法

訓練例集合と外部コーパスでは，そこに含まれる語彙に違いがあるため，蛋白質名そのものは関連文選択の手掛かりにならない．そこで，このような違いの影響を受けずに関連性のある文を抽出するため，パターンを導入する．本論文では，パターンとして，(1) 蛋白質名の前後に現れる単語に基づくパターンと，(2) 蛋白質名を含む構文情報に基づくパターンを提案する．

3.1 前後の単語を基にしたパターン

文中において蛋白質名の前後に頻出する単語を基にパターンを作成し，そのパターンにマッチする文を外部コーパスから抽出することで，訓練例を拡張する．具体的には，蛋白質の前後 1 単語のパターン，前 2 単語又は後 2 単語からなるパターンを作成する．さらに，作成したパターンを訓練例に対してマッチングを行い，一定以上の精度で蛋白質名を抽出できるパターンのみを採用する．前後一単語のパターンの例を次に示す．

“the <keyword> transcription”

“promote <keyword> cell”

前 2 単語又は後 2 単語のパターンの例を以下に示す “activation of <keyword>”

“<keyword> in expression”

3.2 構文情報を基にしたパターン

文献によっては，語彙の違いだけでなく，文の構造が異なる場合があり，訓練例とは異なる構造を持った文を訓練例に加えることで，蛋白質名の判定精度に悪影響を及ぼすことが考えられる．そこで，訓練例の各文を構文解析し，蛋白質を表す単語を含む主語・述語・目的語 (SVO) の関係をパターンとして抽出する．例えば，次のような文について考える．

"Continuous movement at maximal velocity thus requires more than one **kinesin** head." 構文解析の結果，SVO として [movement, require, (kinesin head)] が得られる．ここで，"(kinesin head)" は名詞句を表す．"kinesin" は蛋白質名であるため，結果として "movement require <keyword>" というパターンが得られる．

4 評価実験

コーパスとして，GENIA corpus[7] および BioCreAtIvE 1 の Task 1B のコーパスの内 "fly" と "mouse" の合計 3 種類のデータセットを使用した．同一のコーパスから訓練例・テスト例集合を生成し，他の 2 つのコーパスを外部コーパスとして利用する．前後の単語を基にしたパターンを採用するための閾値は 0.5 とする．拡張後の訓練例集合を基に，蛋白質名判定を行うための分類器として CRF(Conditional Random Fields)[8] を使用した．このとき，各単語を 1 つの事例とし，属性としては品詞・語幹・チャンキングの結果を与える．さらに，CRF には単語を系列として与えるが，その与え方として原文中の単語列のまま与える方法と，主語・述語を中心とする構文解析木の骨格のみを抽出し，その骨格に含まれる単語のみを系列として与える方法を試みた．実験に使用した文数と学習結果をテスト例集合に適用して評価した結果を Table 1 にまとめる．"f", "m", "G" はそれぞれ fly, mouse, GENIA を表し，"f-m" は fly セットから訓練例・テスト例集合を，mouse セットから外部コーパスを生成したことを意味する．さらに，"#", "R", "P", "F" はそれぞれ文の数，再現率，適合率，F 値を表す．なお，各手法を適用した場合の文の数は，追加された文の数を意味する．例えば，"f-m" では，訓練例集合中の文の数 90,309 に対して，前後 1 単語のパターンを適用することで 124 の文を追加し，また外部コーパス全体を追加すると 209,449 の文を追加したことになる．

原文のまま分類器に与えた場合，訓練例に外部コーパス全体を追加すると，全ての場合において F 値が低下しており，訓練集合を増やすことが必ずしも精度向上に貢献するのではなく，むしろ，関連性の低い事例集合による負の影響が確認できる．前後 1 単語・2 単語のパターンを適用した結果，多くの場合で F 値の上昇が見られる．構文パターンを適用した場合では，適合率が低下するものの，再現率の上昇によって F 値も上昇している．提案手法によって抽出される文の数は，全て数百以下である．これらを 10 万前後の文を含む訓練例に追加することによって精度の上昇が見られたことから，蛋白質名判定の精度向上に寄与する文を適切に選択できていることが確認できる．

また，構文解析木の骨格を系列として分類器に与えた場合，構文パターンを適用した場合の精度が，一例を除き，最も良い結果となっている．これは，構文構造を基にしたパターンでは，骨格のみを利用するという点にうまく適応し，その結果として有用な文が抽出できたためと考えられる．一方，構文木の骨格のみを分類に利用するということから，原文での単語の前後関係は重要視されなくなり，その影響から前後 1 単語・2 単語のパターンでは精度の上昇がみられなくなっている．

5 おわりに

本論文では，訓練例が十分ではない場合に，蛋白質名判定の精度を向上させることを目的とし，外部コーパスから有用な文を抽出し訓練例を拡張する手法を提案した．抽出手法としては，蛋白質名の前後に出現する単語を基にしたパターン，構文構造を基にしたパターンを利用する手法を提案し，適用実験を行った．その結果，これらの手法を適用した場合に精度の向上が見られることから，提案手法の有効性を確認した．

今後の課題としては，本論文では各パターンごとに訓練例の拡張を行い比較したが，これらのパターンを組み合わせて，単一のパターンでは誤分類するような事例を補完することで精度向上を図ることが考えられる．

参考文献

1) K. Miyanishi, M. Takeuchi, T. Ozaki and T. Ohkawa. Iterative learning with feature update for extracting sentence containing protein function information. *7th Atlantic Symposium on Computational Biology & Genome Informatics(CBGI2007)*, pp. 96–102, 2007.

2) Md. A. Munna and T. Ohkawa. A method to extract sentences with protein functional information from literature by iterative learning of the corpus. *IPSJ Transactions on Bioinformatics*, 47(SIG 17(TBIO 1)), pp. 22–30, 2006.

Table 1　各事例集合の全組み合わせにおける適用実験結果

		原文のまま						骨格の系列					
		f-m	m-f	f-G	G-f	m-G	G-m	f-m	m-f	f-G	G-f	m-G	G-m
訓練例のみ	#	90,309	105,498	90,309	118,891	105,498	118,891	3,569	4,014	3,569	4,107	4,014	4,107
	R	0.36	0.12	0.36	0.68	0.12	0.70	0.38	0.14	0.38	0.63	0.14	0.61
	P	0.74	0.48	0.74	0.82	0.48	0.82	0.56	0.45	0.56	0.72	0.45	0.78
	F	0.48	0.19	0.48	0.74	0.19	0.76	0.45	0.22	0.45	**0.67**	0.22	0.69
前後1単語	#	124	7	57	10	31	12	118	7	66	6	29	10
	R	0.48	0.12	0.36	0.79	0.12	0.70	0.37	0.14	0.40	0.61	0.14	0.61
	P	0.68	0.48	0.72	0.77	0.46	0.82	0.55	0.45	0.54	0.73	0.45	0.78
	F	**0.56**	0.19	0.48	**0.78**	0.19	0.75	0.44	0.22	0.46	0.66	0.21	0.69
前後2単語	#	266	112	233	46	423	94	254	110	70	41	388	89
	R	0.41	0.12	0.39	0.78	0.16	0.68	0.36	0.15	0.40	0.63	0.17	0.61
	P	0.73	0.46	0.66	0.76	0.34	0.82	0.55	0.45	0.52	0.72	0.35	0.78
	F	0.53	0.19	**0.49**	0.77	0.21	0.75	0.43	0.22	0.45	0.67	0.23	0.68
構文パターン	#	259	331	627	331	721	242	220	281	565	280	648	209
	R	0.42	0.16	0.46	0.71	0.21	0.71	0.46	0.25	0.49	0.62	0.36	0.63
	P	0.59	0.41	0.50	0.74	0.27	0.82	0.51	0.39	0.46	0.70	0.32	0.77
	F	0.49	**0.23**	0.48	0.72	**0.23**	**0.76**	**0.49**	**0.31**	**0.48**	0.66	**0.34**	**0.70**
外部全て追加	#	209,449	182,473	367,369	182,473	486,523	209,449	7,971	7,183	12,706	7,174	16,879	7,983
	R	0.32	0.08	0.32	0.66	0.08	0.69	0.30	0.24	0.42	0.57	0.39	0.42
	P	0.80	0.63	0.80	0.84	0.63	0.84	0.59	0.34	0.35	0.70	0.20	0.82
	F	0.46	0.14	0.46	0.74	0.14	0.76	0.40	0.28	0.38	0.63	0.26	0.55
テスト例	#	181,983	182,473	181,983	367,369	182,473	367,369	7,140	7,770	7,140	12,772	7,770	12,772

3) G.D. Zhou, D. Shen, J. Zhang, J. Su, and S.H. Tan. Recognition of protein/gene names from text using an ensemble of classifiers. *BMC Bioinformatics 2005, 6(Suppl 1):S7*, 2005.

4) M. Murata, T. Mitsumori, and K. Doi. Overfitting in protein name recognition on biomedical literature and method of preventing it through use of transductive SVM. *Proceedings of the International Conference on Information Technology*, pp. 583–588, 2007.

5) A. Koike and T. Takagi. Gene/protein/family name recognition in biomedical literature. *Proceedings of BioLINK 2004: Linking Biological Literature, Ontologies, and Databases*, pp. 9–16, 2004.

6) L. Hirschman, M. Colosimo, A. Morgan, and A. Yeh. Overview of biocreative task 1b: Normalized gene lists. *BMC Bioinformatics 2005, 6(Suppl 1):S11*, 2005.

7) N. Collier, H.S. Park, N. Ogata, Y. Tateishi, C. Nobata, T. Ohta, T. Sekimizu, H. Imai, K. Ibushi, and J. Tsujii. The genia project: corpus-based knowledge acquisition and information extraction from genome research papers. *Proceedings of the ninth Conference on European Chapter of the Association for Computational Linguistics*, pp. 271–272, 1999.

8) J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289, 2001.