情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2009-MPS-76 No.38
Vol.2009-BIO-19 No.38
2009/12/18

# コーパスからの文選択による事例集合拡張に基づく蛋白質名判定

宮西一徳[†1]　尾崎知伸[†2]　大川剛直[†3]

蛋白質構造解析に関する文献の増加に伴い，文献中の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，訓練例が十分ではない場合，高い判定精度が得られない．そこで，利用可能な外部のコーパスによる訓練例集合の拡張を考えるが，単純にコーパスを追加するだけでは，悪影響を受けるおそれがある．従って，本論文ではコーパスから精度向上に有効な文を選択する手法を提案する．評価実験では，提案手法によって，訓練例が少ない場合に判定精度の向上が確認できた．

# Identification of Protein Names Based on Extending an Instance Set by Selecting Sentences

Kazunori Miyanishi,[†1] Tomonobu Ozaki[†2] and Takenao Ohkawa[†3]

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. Then we propose a method to select sentences from a corpus, which are effective for identification. In the experiment, it was confirmed that the accuracy was improved by the proposed method especially when a training data set is small.

†1 神戸大学大学院自然科学研究科
Graduate School of Science and Technology, Kobe University
†2 神戸大学自然科学系先端融合研究環
Organization of Advanced Science and Technology, Kobe University
†3 神戸大学大学院工学研究科
Graduate School of Engineering, Kobe University

## 1. はじめに

蛋白質の機能情報は，新薬の開発や生命現象の解明など様々な分野において有用である．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いるため，文献から機能情報を自動的に抽出することが求められている．そこで，機能情報抽出の前処理として，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利用して，蛋白質名を特定する手法に関するものである[3)–5)]．このような手法は，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．この問題に対して，蛋白質名に対してタグ付けされているコーパスを利用し，訓練例を拡張することによって判定精度向上を図ることを考える．ここで，異なる生物種に関する文献間では，蛋白質名や遺伝子名などの語彙が異なるため，訓練例とは異なる生物種を対象とする文を追加すると，判定精度に悪影響を及ぼす可能性がある．従って，単純にコーパス全体を訓練例に追加するのではなく，コーパスから判定精度に良い影響を及ぼす文を選択する必要がある．そこで，本論文では，コーパスから判定精度向上に有効な文を選択し，訓練用の事例集合を拡張する手法を提案する．事例集合拡張の流れを図 1 に示す． まず，対象とする文献の一部の文には，あらかじめ蛋白質名に対してタグ付けがなされているものとし (1)，残りの文にはタグ付けがされておらず (2)，最終的にはこのタグ付けされていない文集合に対して蛋白質名判定を行いたいものとする．コーパスから判定精度に有効な文を選択し (3)，対象文献中のタグ付けされた文集合に追加することで，事例集合の拡張を行う．

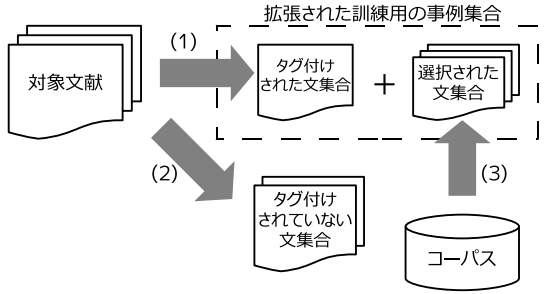


**図 1** Concept of training data extension

Vol.2009-MPS-76 No.38
Vol.2009-BIO-19 No.38
2009/12/18

上記の通り，対象とする文献とコーパスでは，含まれる語彙に違いがある可能性があるため，蛋白質名など特定のキーワードに基づく文選択よりも，文の構造に基づく文選択の方が効果的であると考える．そこで，判定精度向上のために有効な文の選択手法として，文の構造上の特徴に対して重みを設定し，その重みを繰り返し更新することによって効果的な文を選択する手法を提案する．さらに，提案手法により選択された文を，対象文献中のタグ付けされた文集合に追加し，それを基にタグ付けモデルを学習する．このモデルを対象文献中のタグ付けされていない文集合に適用することで，提案手法の有効性を評価する．

## 2. コーパスからの文の選択手法

文献は，対象とする生物種によって特有な語彙（蛋白質名や化合物名など）を持つ傾向がある．従って，対象とする文献とは異なる文献集合であるコーパスから有効な文を選択しようとする場合，蛋白質名自体は手掛かりにならないと考えられる．そこで，蛋白質名を含む文の構造に着目する．蛋白質名を含む文の抜粋を図 2 に示す．ここで，各文献（PMID:10381570 と PMID:10455134）が対象としている生物種は，それぞれ fly と human である．“*Ttk*” と “*AML1 and BSAD*” が蛋白質名であり，主語・述語・目的語の組はそれぞれ “*Ttk* activates transcription” と “*AML1 and BSAD* activate transcription” である．この 2 つの文の間では，蛋白質名自体は異なるが，蛋白質名に関係する述語と目的語は共通である．以下では，このような構造に基づく文の選択手法について述べる．

### 2.1 蛋白質名を含む文

文中での蛋白質名との係り受け関係や共起関係を，文の構造的な特徴として考える．蛋白質名を含む文の特徴的な構造は，構文解析の結果から得られる．ここで，構文解析器として the Stanford parser[6] を使用した．図 2 において，主語・述語・目的語を文の構成要素と考え，主語–述語の関係である “[protein] activates” が文の構造的な特徴として得られる．同様にして，係り受け関係（主語–述語，述語–目的語，前置詞で繋がる名詞ペア）や共起関係（文中で蛋白質名と同時に出現する語）を，各文の特徴として割り当てる．

*... Ttk protein strongly activates transcription, ...* (PMID:10381570)
Subject Verb Object

*... AML1 and BSAD synergistically activate blk promoter transcription ...*
Subject Verb Object

(PMID:10455134)

**図 2** 蛋白質名を含む文

### 2.2 文の構造に基づく選択手法

どのような構造を持つ文が蛋白質名判定に有効であるかということは，事前に明らかではない．従って，判定精度向上に寄与する文を選択する上で手掛かりとなる，構造上の特徴を発見する必要がある．そこで，タグ付けされた文集合を妥当性チェック用集合と訓練例集合に分割する．さらに，文の特徴に重みを設定し，妥当性チェック用集合に対する判定精度を向上させるように重み更新を繰り返す処理を行う．この処理によって，どのような特徴を持った文が，判定精度の向上に寄与するのか特定することができる．ここで，タグ付けモデルの詳細については次節で述べる．文選択のための繰り返し処理の概要を図 3 に示す． 最初に，タグ付けされた文集合を妥当性チェック用集合と訓練例集合に分け（①），コーパスからはランダムに文を選択する（②）．次に，訓練例集合とコーパスから選択された文集合から，タグ付けモデルを学習し妥当性チェック用集合に適用する（③）．妥当性チェック用集合に対するタグ付け結果に基づいて，文の特徴に対する重みを更新する（④）．以降の繰り返し処理では，更新された重みに基づいて文を選択し，訓練例拡張を行う．タグ付け精度が上昇する間，以上の処理を繰り返す．

特徴の重み更新の処理を図 4 に示す．ここで，$F_j(1 \leq j \leq M)$ はコーパス中の文の属性，$M$ は全ての属性の数を表す．また，$f_i(1 \leq i \leq n)$ はタグ付けを誤った単語を含む文の属性，$n$ は属性の数を表す．$W_{F_j}$ は属性 $F_j$ の重みを表し，$W_{F_j}$ の初期値は，$F_j$ の訓練例集合中での出現回数により正規化した値である．さらに，$U(> 0)$ は，重みの増加率である．この

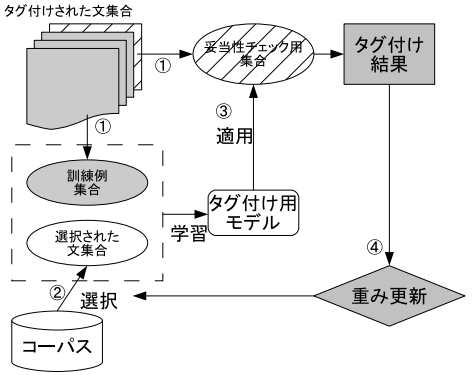


**図 3** 文選択の概要

Vol.2009-MPS-76 No.38
Vol.2009-BIO-19 No.38
2009/12/18

**Procedure** : update weights of features
**1 for** $(i = 1..n)$
**2** **for**$(j = 1..M)$
**3** **if**$(F_j == f_i)$
**4** $W_{F_j} = W_{F_j} \times U$

**図 4** 文の特徴に対する重みの更新処理

**Procedure** : update weights of sentences
**1 initialize** : $\tilde{W}_S = 0$
**2 for**$(i = 1..N)$
**3** **for**$(j = 1..M)$
**4** **if**$(S_i \; has \; the \; feature \; F_j)$
**5** $\tilde{W}_{S_i} = \tilde{W}_{S_i} + W_{F_j}$
**6 for**$(i = 1..N)$
**7** $W_{S_i} = \tilde{W}_{S_i} \Big/ \sum_{k=1}^{N} \tilde{W}_{S_k}$

**図 5** 文の重み更新処理

処理によって，タグ付けを誤った単語を含む文の属性の重みは，$U$ 倍に増加する． 次に，文の重み更新の処理を図 5 に示す．ここで，$S_i(1 \leq i \leq N)$ はコーパス中の文，$N$ は全ての文の数，$W_{S_i}$ は文 $S_i$ の新たな重みを表す． 最終的に，新たな重み $W_{S_i}$ は，**7** 行目で正規化された値として得られる．最後に，更新された重みに基づいて文を選択する処理を図 6 に示す．ここで，$W_{S_i}$ は文 $S_i$ の新たな重み，$W'_{S_i}$ は更新前の重みを表し，関数 $rank(W)$ は全ての文の重みの中での重み $W$ のランクを返す． 結果として，更新によって重みが増加した文，または全ての文の中で重みにおいて上位 $T_R$ に入る文を選択する．

## 3. タグ付けモデルと蛋白質名判定

本論文では，蛋白質名のタグ付けに CRF(Conditional Random Fields)[7),8)] を使用する．前処理として，Brill's tagger[9)] により品詞タグ付け，CRF++[10)] によりチャンキングを行う．入力された文の各単語の属性として，ステミング・品詞・チャンキングを使用する．さらに，ある単語のタグ付けを行う際に，前後 2 単語の属性を利用する．タグ付けの例を図 7 に示す．

**Procedure** : select sentences
**1** select $= \{\}$
**2 for**$(i = 1..N)$
**3** **if**$(W_{S_i} > W'_{S_i} \parallel rank(W_{S_i}) \; is \; superior \; to \; T_R)$
**4** select $=$ select $\cup \{S_i\}$

**図 6** 文選択処理

| word | stem | pos | chunk | tag |
|---|---|---|---|---|
| at | at | IN | B-PP | O |
| position | position | NN | B-NP | O |
| 187 | 187 | CD | I-NP | O |
| in | in | IN | B-PP | O |
| esterase | esterase | NN | B-NP | K |
| 6 | 6 | CD | I-NP | K |
| contributes | contribute | VBZ | B-VP | O |
| significantly | significantly | RB | B-ADVP | O |

**図 7** “esterase” に対してタグ付けする場合の例

前節で述べた通り，タグ付けされた文集合は，訓練例集合と妥当性チェック用集合に分割される．従って，文集合の分け方によって，複数のタグ付けモデルが得られ，同じ数のタグ付け結果が得られる．これらの複数のタグ付け結果を組み合わせることによって，最終的なタグ付け結果を得る（図 8）． ここで，閾値 $T_M$ を導入する．つまり，ある単語について，$T_M$ 個以上のモデルが蛋白質名であると予測しているとき，最終的にその単語は蛋白質名であると判定する．

## 4. 評　　価

GENIA corpus[11)] と BioCreAtIvE1 Task 1B[12)] の “mouse” と “fly” のアブストラクト集合を使用して，提案手法の有効性を評価する．GENIA corpus 中の数百アブストラクトを入力文献として使用する．そのうち，10 から 100 アブストラクトをタグ付けされた文集合とし，500 アブストラクトを評価用の文集合とする．“mouse” と “fly” の 5000 アブス

Vol.2009-MPS-76 No.38
Vol.2009-BIO-19 No.38
2009/12/18

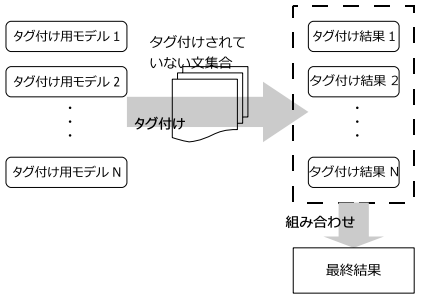


**図 8** タグ付け結果の組み合わせ

トラクトをコーパスとして使用する．さらに，タグ付けされた文集合のうち 90%を訓練例集合とし，残りを妥当性チェック用集合とする．ここで，10 通りの文集合の分け方を試みる．従って 10 種類のタグ付けモデルが得られる．重み更新の繰り返し処理において，初回にコーパスから 800 文をランダムに選択し，訓練例集合に追加する．各ステップにおける属性の重み増加率 $U$ は 2，閾値 $T_R$ は 1000 とする．さらに，重みが増えた文のうち上位 500 文を選択する．各タグ付けモデルは弱学習器ではないため，単純な多数決による結果が最も良い精度になるとは限らない．従って，予備実験の結果に基づいて，タグ付けモデルの組み合わせの閾値 $T_M$ は 1 とする．

提案手法を以下の 3 つのベースラインと比較する．**Baseline (1)**: タグ付けされた文集合のみで学習したモデルを使用（コーパスは不使用）．**Baseline (2)**: タグ付けされた文集合とコーパス全体を使用してモデルを学習．**Baseline (3)**: タグ付けされた文集合にコーパスからランダムに選択した 50 アブストラクトを追加し学習．F 値による各手法の比較を図 9 に示す． 各ベースラインの結果から，コーパス全体を訓練例集合に追加した場合に大きく精度が低下しており，ランダムに 50 アブストラクトを追加した場合でも，訓練例集合のみで学習した場合に比べて精度が低下していることが分かる．一方，提案手法は，十分な訓練例が与えられている場合には，訓練例のみの場合とほぼ同じ精度であるが，訓練例が少ない場合に精度が高い．特に訓練例集合が 30 アブストラクト以下の場合，すべてのベースラインの精度が大きく低下しているのに対して，提案手法はほとんど低下していない．次に，訓練例集合の文の数と，Baseline (3) と提案手法によって訓練例集合に追加された文の数の平均を図 10， 11 に示す． ここで，“mouse” と “fly” の各コーパス中の文の総数は，それぞれ 41,345，38,510 である．このコーパス中の文の数が，訓練例集合中の文の数に比

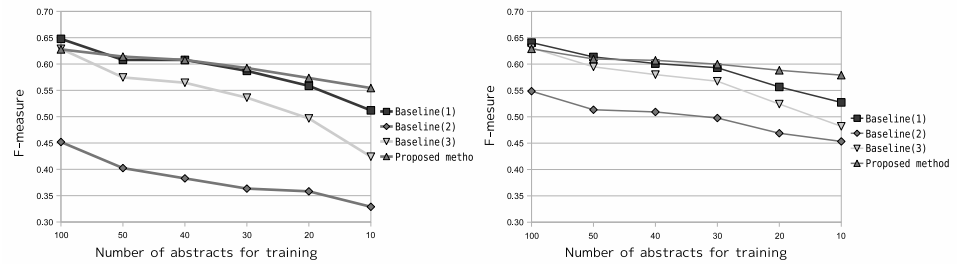


**図 9** 各手法の精度比較 (“mouse” （左），“fly” （右）)

べて非常に多いため，コーパス全体を訓練例集合に追加すると，異なる語彙に基づく悪影響を受けている．さらに，ランダムに 50 アブストラクトを追加した場合，追加される文の数は約 400 と少ないにもかかわらず，精度に悪影響を及ぼしている．一方，提案手法では，1,200 文から 1,400 文を追加しているが，特に訓練例が少ない場合に精度が向上している．このことから，提案手法によって，少ない訓練例を補完し精度向上に有効な文を選択できていることがわかる．

次に，各ベースラインと提案手法の再現率，適合率，F 値の比較を行う．タグ付けされた文集合が 10 アブストラクトの場合における各手法の再現率，適合率，F 値を図 12 に示す． コーパス全体を追加した場合 (Baseline (2)) が，再現率が最も低く，適合率が最も高い．これは，タグ付けモデルを学習するのに多くのデータを利用することが，適合率に対して良い影響を及ぼしていると考えられる．しかしながら，再現率には悪影響を及ぼして

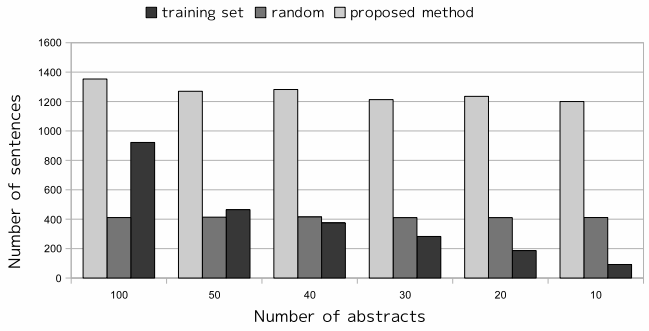


**図 10** 訓練例集合の文の数と追加された文の数 (“mouse”)

Vol.2009-MPS-76 No.38
Vol.2009-BIO-19 No.38
2009/12/18

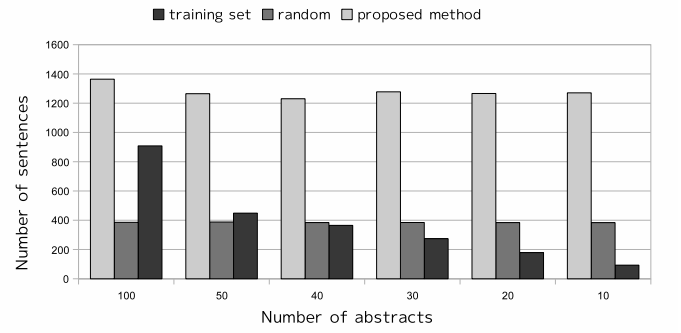


**図 11** 訓練例集合の文の数と追加された文の数 (“fly”)

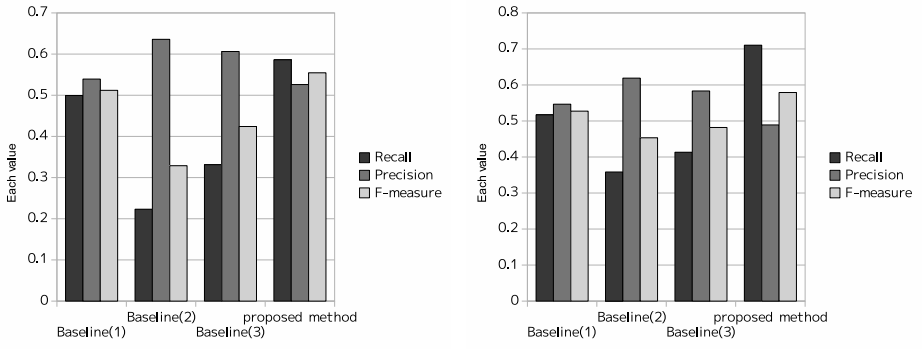


**図 12** 再現率, 適合率, and F 値 (“mouse” (左), “fly” (右))

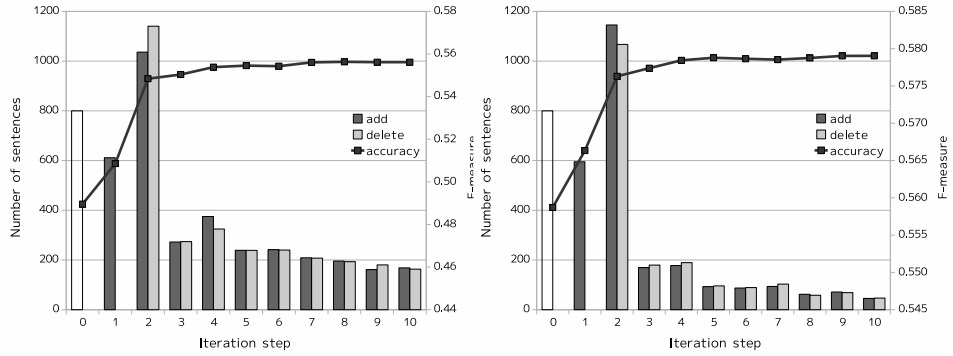


**図 13** 追加・削除される文の数と精度 (“mouse” (左), “fly” (右))

おり，結果として F 値は最も低くなっている．Baseline (3) の結果も同様の傾向を示している．しかし，提案手法は，適合率は Baseline (1) より少し低下しているが，再現率が最も高い値となっている．この結果は，提案手法がコーパスによる悪影響を受けることなく，適切な文を選択できているということを示している．

最後に，重み更新の繰り返し処理において，追加・削除される文の数と判定精度の推移について考察する．タグ付けされた文集合が 10 アブストラクトの場合における，繰り返し処理の各ステップでの追加・削除される文の数の平均と精度を図 13 に示す． 最初に追加される文の数は，前述の通り 800 であり，この文の数が少ないため，1 回目の繰り返しステップでは，文の追加のみが行われる．2 回目のステップで 1,000 以上の文が追加・削除され，それ以降は大きく減少する．精度は，最初の 2 回のステップで上昇し，それ以降は収束する．従って，繰り返し処理の中で，蛋白質名判定に有効な文が選択され，判定精度に悪影響を及ぼす文は削除されていることがわかる．

## 5. おわりに

本論文では，蛋白質名判定に有効な文をコーパスから選択し，訓練例拡張を行う手法を提案した．提案手法では，あらかじめタグ付けされた文集合を訓練例集合と妥当性チェック用集合に分割し，文の構造上の特徴に基づいた文選択を繰り返し行うことで，最適な文選択に基づく訓練例拡張を行う．評価実験の結果，提案手法による精度が，コーパスを使用しない場合やコーパスを使用した単純な訓練例拡張による精度に対して高くなることが確認され

Vol.2009-MPS-76 No.38
Vol.2009-BIO-19 No.38
2009/12/18

た．特に，訓練例が少ない場合に，判定精度に顕著な差が出ることが明らかとなった．以上のことから，提案手法により，蛋白質名判定に有効な文が選択できることが示された．

今後の課題としては，文選択に利用した文の構造上の特徴を，タグ付けモデルの学習にも利用することで，モデルの判定制度の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. In *Proceedings of the 7th Atlantic Symposium on Computational Biology & Genome Informatics (CBGI2007)*, pp. 96–102, Salt Lake, USA, 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.

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. In *Proceedings of the 10th International Conference on Information Technology*, pp. 583–588, Rourkela, India, 2007.

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

6) D.Klein and C.D.Manning. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423–430, Sapporo, Japan, 2003.

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

8) F.Sha and F.Pereira. Shallow Parsing with Conditional Random Fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 134–141, Edmonton, Canada, 2003.

9) E.Brill. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, Vol. 21, pp. 543–565, 1995.

10) T.Kudo. CRF++: Yet Another CRF toolkit. Available at *http://crfpp.sourceforge.net/*, 2005.

11) 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. In *Proceedings of the Ninth Conference on European chapter of the Association for Computational Linguistics*, pp. 271–272. Association for Computational Linguistics Morristown, NJ, USA, 1999.

12) 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.