2005－FI－81（5）
2005／11／2

ライフサイエンス分野向けテキストマイニング用辞書における不適切エントリーの抽出

竹内広宜* 吉田一星* 伊川洋平* 飯田一雄† 福井要子†

*日本 IBM 株式会社 東京基礎研究所
†東北化学薬品株式会社 生命システム情報研究所

*{hironori,issei,yikawa}@jp.ibm.com
†{iida,fukui}@t-kagaku.co.jp

テキストマイニングにおける単語の出現頻度の集計では、表記のゆれを吸収するため辞書が用いられる。その際、ライフサイエンス分野では外部リソースを用いてこのテキストマイニング用辞書を構築することが多いが、元リソースの質などが原因で単語の出現頻度集計に悪影響を与える不適切なエントリーが多く存在している。本研究では、構築された辞書から不適切な辞書エントリーを抽出しランキングする方法を提案し検証した。

キーワード:テキストマイニング、辞書

Detection of Invalid Entries in the Dictionary of Text Mining for Biomedical Documents

Hironori TAKEUCHI* Issei YOSHIDA* Yohei IKAWA* Kazuo IIDA† Yoko FUKUI†

*IBM Research, Tokyo Research Laboratory, IBM Japan, Ltd.
†Research Institute of Bio-system Informatics, Tohoku Chemical Co., Ltd.

In the calculation of keyword frequency distribution in the text mining, we usually use a dictionary to map different keyowrds to a canonical form. In the biomedical domain, we construct the dictionary from external resources. However there are many invalid entries that make a negative impact on the calculation of keyword frequency. In this paper, we propose the methods to extract invalid entries from the dictionary and examine their effectiveness.

Keywords:Text Mining, Dictionary

1　はじめに

膨大なテキストデータの蓄積とともに、有用な情報の抽出や傾向の分析を行うためのテキストマイニング技術が研究開発され、様々な分野で用いられはじめている [5]。特にライフサイエンス分野では MEDLINE と呼ばれる論文の書誌情報と要約を収録した公共のデータベースがあり検索ツール PubMed を用いて自由にアクセスすることが可能となっている。MEDLINE はライフサイエンス分野のほぼ全ての論文誌を網羅しており、1400 万文書を越えるその膨大なリソースは自由に用いることができる。そのため MEDLINE からの情報抽出やその有効活用を目的としたテキストマイニング技術が近年注目をあびている [8][10]。

通常、テキストマイニングでは主に自然言語処理を中心とした技術を用いて前処理と呼ばれる情報抽出を行う。この前処理で得られた情報を用いて、多くのテキストマイニングシステムでは特定の文書集合中でのキーワードの出現頻度の分布などを求め、傾向分析などに用いている。前処理では自然言語処

理で得られた結果から、遺伝子やタンパク質などの属性抽出や係り受けをもとにした関係情報を抽出するが、この過程において辞書を用いて表記を統一化する処理が行われる。これは、同じ事柄を指す言葉に表記のゆれがある場合に、これらの表記語（同義語とも呼ばれる）をあるひとつの見出し語に置き換える処理 (以降、異表記解消処理と呼ぶ) である。特にライフサイエンス分野においては、正式な学術名だけでなく略称や研究コミュニティ独自の表記などが多く存在するため、ある特定の事柄を表す単語の出現頻度をより正確に把握するためにはこの異表記解消処理が必要不可欠となっている。

異表記解消処理では見出し語とそれに結びつけられた複数の表記語のペアが蓄えられた辞書 (同義語辞書とも呼ばれる) を用いる。文書中に出現した単語についてこの辞書を適用し、その単語が表記語として辞書にあれば該当する見出し語に置き換える処理を行う。ライフサイエンス分野では遺伝子名などにおいて大文字・小文字の違いで意味が異なる場合があるそのため、文書中の単語と表記語のマッチングではを大文字・小文字の違いを区別することが多い。

本処理で用いる辞書を構築するためには見出し語とその異表記語を収集する必要がある。ライフサイエンス分野では遺伝子、タンパク質などのカテゴリーを中心に様々な外部リソースがあり、そこから見出し語（正式名称）とともに異表記語についての情報も得られることが多い。辞書構築においては複数の外部リソースから辞書エントリーを抽出し各カテゴリーごとに辞書を作成し、最終的に一つの巨大な辞書に統合する。またリソースから得られる異表記の情報は不十分であるため、それを補うためコーパスを処理をして自動的に見出し語と表記語のペア情報を抽出し辞書に追加する方法も提案されている [4]。

しかしながら、外部リソースを用いて構築した辞書の中には不適切な辞書エントリーが多く含まれている。これは主に元リソースの質によるものだが、ある見出し語に結び付けられた表記語群の中に一般名詞を含む高頻度語や異なるカテゴリの見出し語が入っていることがある。したがって、このような状態の辞書を適用すると、例えば文書中に大量に出てくるある一般語がすべて辞書中に登録されている専門用語に置き換えられてしまい、単語の出現頻度の計算において間違った結果を導き出してしまう可能性がある。このような不適切な辞書エントリーと思われる語を除去する方法として一般語辞書を用いて誤って表記語として辞書に登録された一般語を取り除く方法が提案されている [3][9]。しかしながら、不適切と判断される辞書エントリーは一般語だけではなく、また同じライフサイエンス分野においても対象や前処理結果の活用方法などによって変わるものもある。したがって、その度に辞書管理者が膨大なエントリーを調査し目的に応じた辞書に修正することが理想であるが、これは非常に困難な作業である。

本研究では、辞書エントリーの中から不適切と思われるエントリーを不適切度を示すスコアとともに抽出する手法について述べる。上述の通り、辞書中のエントリーには目的に応じて除去すべきかどうかの判断が変わる（以降、グレーゾーンと呼ぶ）エントリーが多く、最終的には専門家が判断する必要がある。このような状況において、辞書エントリーが不適切度スコア順に並んでいれば専門家による判断は容易になり、辞書のメンテナンスに要するコストを下げる効果があると考えられる。本章以降の構成は次のようになっている。2 章で本研究で扱う辞書とその問題点について述べ、3 章で不適切エントリーの抽出方法について述べる。4 章で実際のデータを使った実験について述べた後、5 章で結果・考察を行い、最後にまとめを行う。

2 ライフサイエンス分野向けテキストマイニングにおける辞書

2.1 複数の外部リソースを用いた異表記解消処理用辞書

ライフサイエンス分野においては、実験などで得られた遺伝子、タンパク質などの情報が様々なデータベースで公開されている。例えば Entrez Gene(旧 LocusLink[6]) には各遺伝子についての配列情報をはじめ様々な情報がある。このデータベースでは各

遺伝子には Official Symbol と呼ばれる名称が与えられ、別称が Gene Aliases や Other Aliases として与えられている。このような情報から Official Symbol を見出し語、Gene Aliases および Other Aliases を表記語とした辞書エントリーを作成することができる。同様なデータベースがタンパク質についても存在する (例えば SwissProt[1])。また、ライフサイエンス分野に関係した術語を集めたリソースとして UMLS[2] がある。

このような複数の外部リソースから見出し語-表記語のペアとして定義される辞書エントリーを収集し、各カテゴリー (遺伝子、タンパク質など) ごとに辞書を作成する。そして最終的にこれらの辞書を統合することでライフサイエンス分野向けの巨大な言語処理用辞書を構築することができる。ライフサイエンス分野は専門用語が多く表記も多岐にわたるため、既存の専門データベースを利用して言語処理用辞書およびオントロジーを構築することは有効であり、広く行われている。

テキストマイニングの前処理では、文中に出てくる単語が辞書中の表記語である場合には紐付けられている見出し語に置き換える異表記解消処理を行う。文中に出てくる単語の中には、ほぼ辞書中にある表記語と同じであるものの、微妙な表記の違い (例えばハイフンや空白の有無) があるため辞書エントリーとの完全一致では処理しきれない (低再現率) 場合があることが報告されている [4]。このような問題に対しては、学習データを用いて、辞書にない細かい表記のゆれを考慮した対象語を動的に構成し、再現率を上げる方法が提案されている [7]。

2.2 辞書中の不適切エントリーがテキストマイニングに与える問題点

通常、人手による専門用語辞書の構築はコストと時間がかかるため、複数の外部リソースから辞書エントリーを収集し、巨大な辞書を構築することは非常に有用である。しかしながら、収集された膨大な辞書エントリーの中には不適切なエントリーが入ることが多い。例えば、前節のように複数の外部リソースから作成した辞書には一般語として使われる言葉が、ある辞書エントリーの見出し語に対応する表記語の中に入る場合がある。これは主に辞書構築の元になった外部リソースの質に原因がある。例えば、Entrez Gene で Spna2(Gene ID:20740) を検索すると、Other Aliases として brain が登録 (2005 年 7 月現在) されており、このデータから辞書エントリーを収集してしまうと遺伝子見出し語”Spna2”の表記語として”brain”が登録されてしまう。データの登録ミスや機械的にデータ収集してリソースを作成したことなどが、外部リソースの質を下げる要因として挙げられる。

このような不適切な語が辞書の表記語として存在すると単語の出現頻度の計算などに悪影響を与える可能性がある。例えば、一般語が辞書のある見出し語の表記語として登録されてしまうと、テキストマイニング前処理における異表記解消処理において大量に出現する一般語が対応する見出し語に置き換えられてしまう。その結果、本来は高頻出ではない遺伝子やタンパク質の見出し語を頻度計算において高頻度語として抽出してしまい、間違った知見を導き出してしまう危険性がある。また、複数のリソースから得られたエントリーを統合するためあるカテゴリーの見出し語に紐つけられた表記語が別のカテゴリーの見出し語となっている場合がある。この場合も、異表記解消処理において異なる分野の言葉が対応する見出し語に置き換えられてしまい、不適切な結果を導き出す可能性がある。

高頻度の一般語を抽出する方法としては、WordNet のようなインターネット上で公開されている一般語辞書を用いる方法がある [3][9]。この方法によって誤って専門用語である見出し語の表記語として辞書に登録されてしまった高頻度の一般語を除去することができるが、一般語辞書にはない不適切辞書エントリーを抽出することは難しい。本研究では辞書に登録されたエントリー (見出し語および表記語) の特性に注目し、不適切なエントリーを不適切度を表すスコアとともに抽出する手法を提案する。

3 不適切辞書エントリーの抽出方法

3.1 不適切辞書エントリーとリスクスコア

本研究ではキーワードの出現分布に悪影響が出る不適切辞書エントリーの抽出を目的にしている。そのため、ある見出し語の表記語が表記上不適切であっても、その出現頻度が非常に少なければ見出し語の出現分布への影響は少ない。そこで本研究では、専門家による評価だけでなく、表記語の出現頻度も考慮に入れ、もし表記語が不適切であった場合に見出し語の出現分布にどの程度の悪影響を与えるかというリスクスコアを各正解辞書エントリーに付与する。そして、ある値以上のリスクスコアを持ったエントリーを不適切辞書エントリーと定義する。

正解辞書エントリーの表記語 S に対するリスクスコア (rs) は次のように計算する。S の出現頻度 (本研究では PubMed を用いた検索で得る) を f_S とする。また、$[0,1]$ の区間 (0:適切、1:不適切) に正規化された評価者による評価値を ν とし、以下のようにリスクスコアを定義する。

$$rs(S) = \left(\frac{\log(f_S + 1)}{\log(f_{max} + 1) \cdot n} \nu \right)^a \qquad (1)$$

ここで、f_{max} は対象文書集合中における単語の出現頻度の最大値となる。PubMed を用いて f_S を得る場合には、PubMed を用いた検索で求まる単語の最大出現頻度となるが、$f_{max} = 10^7$ とする。また、a はスコアの立ち上がりを決定するパラメータであるが、0.5 とした。こうして得られたリスクスコアのうち本研究では $rs(S) \geq 0.4$ となる表記語 S を不適切辞書エントリーとした。

3.2 語の出現分布を用いた不適切エントリーの抽出

不適切な辞書エントリーの一つに、本来出現頻度が少ない専門用語である見出し語の表記語として一般語を含む高頻度語が定義されている場合がある。このような場合、キーワードの出現頻度分布の計算において本来高頻度でないキーワードが高頻度語として上位に現れるという悪影響がある。しかしながら、高頻度語の定義というのは分野、カテゴリー、または見出し語によって異なるので、ある閾値以上の頻度を持つものを高頻度語として不適切辞書エントリーとして抽出することはできない。そこで、本研究では見出し語の出現頻度を考慮し、それに紐付けられている表記語が高頻度語かどうかを判定することを考える。

文書数 N のテストコーパスにおける、見出し語 c_i の頻度を f_{c_i}、表記語 s_j の出現頻度を f_{s_j} とする。このとき、見出し語 c_i および表記語 s_j の出現確率の推定値は $\hat{p}_{c_i} = \frac{f_{c_i}}{N}$、$\hat{p}_{s_i} = \frac{f_{s_j}}{N}$ となる。この推定値を用いて、表記語 s_j の出現分布の見出し語 c_i の出現分布に対する Kullback-Leibler 距離 (KL 距離) を以下のように求める。

$$KL(\hat{\boldsymbol{p}}_{s_j}, \hat{\boldsymbol{p}}_{c_i}) = E(\hat{\boldsymbol{p}}_{s_j} \log \frac{\hat{\boldsymbol{p}}_{s_j}}{\hat{\boldsymbol{p}}_{c_i}}) \qquad (2)$$

この KL 距離を用いて、見出し語に紐付けられている表記語の評価を行う。KL 距離の大きい順に辞書エントリーを評価し、ある閾値 τ 以上のもの (本研究では $\tau = 10^{-4}$ とした) を不適切辞書エントリーとして抽出する。テストコーパスはマイニング対象と同質の文書集合を使用するのが適切と考え、本研究では MEDLINE 文書を用いた。

出現確率の KL 距離で評価した場合、見出し語と表記語が同程度の頻度を持つ不適切辞書エントリーを抽出することができない。また、不適切辞書エントリーとして抽出された表記語の中には、実際には見出し語より研究者間では広く知られ、見出し語よりも高頻度で出現するものもある。こういった場合には KL 距離を用いた評価は不十分である。この問題を回避する方法を次節で述べる。

3.3 Web による検索結果を用いた判定方法

見出し語と表記語が同程度の頻度を持つ不適切辞書エントリーについては、頻度情報をもとに検出することはできない。例えば、PTPN1(見出し語)、

Protein-tyrosine phosphatase(表記語) は予備実験で用いたテストコーパス中で同程度の出現頻度を持つ。このように頻度が同程度であった場合、不適切な表記語をそのまま辞書に用いると対応する見出し語の出現頻度は 2 倍以上になってしまい、間違った知見を導く可能性がある。逆に研究者間において表記語が見出し語と同程度または頻繁に用いられる場合、出現頻度分布の比較をもとに評価を行うと、適切な表記語を不適切辞書エントリーと判断し辞書から除いてしまい、有用な情報を多く失ってしまう。

本研究では、文書中で同時に出現する見出し語および表記語の出現パターンを用いて、その辞書エントリーの妥当性を評価する。表記語は見出し語の言い換えであると考えることができるので、見出し語 (canonical)、表記語 (surface) に対して適切な辞書エントリーであれば、"... canonical (surface) ..."といった言い換え表現が文書中にあることが期待される。そこで、語の出現頻度分布をもとに抽出された適切・不適切辞書エントリーに対して見出し語および表記語を含む文書を検索し、以下のように妥当性を再評価する。

見出し語・表記語の出現分布から適切と判断された辞書エントリーについて、検索で得られた文書集合中に言い換えと推測される表現があるかどうかを判断する。表現がなかった場合には不適切エントリーだと推定する。逆に見出し語・表記語の出現分布から不適切と判断された辞書エントリーについては、検索で得られた文書集合中に言い換えと推測される表現があった場合に適切エントリーだと推定する。前者で用いる表現パターンを P1、後者で用いるパターンを P2 とする。それぞれの表現パターンは以下のように定義する。

$$\begin{aligned}
\text{P1}: & \mathit{canonical}(\mathit{surface}), \mathit{surface}(\mathit{canonical}), \\
& \mathit{canonical}/\mathit{surface}, \mathit{surface}/\mathit{canonical}, \\
& \mathit{canonical} - \mathit{surface}, \mathit{surface} - \mathit{canonical} \\
\text{P2}: & \mathit{canonical}(\mathit{surface})
\end{aligned}$$

適切辞書エントリーを判定するために用いられるパターン P1 が多いのは、あらゆる言い換え表現を網羅的に調べ、言い換え表現がないことを推定し不適切エントリーとして抽出する必要があるからである。逆に不適切辞書エントリーを判定する際には、確実な言い換え表現があった場合にのみ適切エントリーとして抽出する必要があるため特定のパターンのみを用いている。

言い換え表現パターンを抽出するためには見出し語および表記語を検索して得られる文書集合が必要である。ライフサイエンス向けテキストマイニングでは主に MEDLINE のような最新の研究成果の概要が対象となっている。MEDLINE は 1400 万文書を越える膨大なライフサイエンス分野の論文概要のデータベースであり、PubMed を通して柔軟な検索を行うことができる。しかしながら、MEDLINE には研究結果の要約であり字数に制約あるため省略形とその正式名称といった記述はあるが、それ以外の異表記に関する言い換えの記述は非常に少ない。そのため、言い換え表現の知識獲得に用いるのは難しい。そこで本研究では一般の Web 検索を用いた。

見出し語・表記語の出現分布から適切・不適切と判断された辞書エントリーについて、それぞれ検索結果文書集合中に該当パターンがみつかった場合、そのエントリーは適切・不適切か判断がつかないグレーゾーンのエントリーと考えることができる。そこで人手評価 ν に相当する値として $\nu(= 0.45)$ を与え、ν および表記語 S の出現頻度 f_S からリスクスコアを推定し、閾値を越えるエントリーを不適切エントリーとし再分類した。

4 実験

4.1 実験に用いた辞書エントリー

本研究では、Entrez Gene, UMLS, Swiss-Prot から作成した同義語辞書の中から見出し語が Entrez Gene の Official Symbol であるもの 200 語とそれに紐づけられている表記語を実験用辞書エントリーとして用いた。見出し語-表記語のペアで構成される辞書エントリーの数は 1137 であった。この辞書エントリーに対し 4 人の専門家が 3 段階 (1：適切、2：場合によって不適切、3：不適切) の評価を行った。

表 1 に評価結果の例を示す。

表 1: 辞書エントリーの評価例

見出し語	表記語	評価 1	評価 2	評価 3	評価 4
Spna2	brain	3	3	3	3
Dhfr2	Dihydrofolate reductase	2	1	2	2
STMN1	leukemia-associated gene	3	3	1	2

本研究では、評価実験で用いるデータに対して複数の専門家 (n 人) による評価値があるため、[0;1] に正規化した j 番目の評価者評価値 s_j から平均評価を用いて $\nu = \frac{1}{n}\sum_i^n s_j$ とし、3.1 で定義したリスクスコアを求めた。表 2 に表 1 と同じ辞書エントリーのリスクスコアを示す。

表 2: リスクスコアの例

見出し語	表記語	ν	f_S	$rs(S)$
Spna2	brain	1.000	645985	0.984
Dhfr2	Dihydrofolate reductase	0.375	4170	0.476
STMN1	leukemia-associated gene	0.625	13	0.346

上記の方法でリスクスコアを付与された正解データに対して、次節で述べるシステムを用いて不適切辞書エントリーの抽出実験を行った。

4.2 不適切辞書エントリー抽出システム

3 章で述べた手法を用いて以下の手順で不適切辞書エントリーの抽出実験を行い評価を行う。

- 見出し語・表記語それぞれについてテストコーパス中での出現頻度を求め、出現頻度分布の比較から不適切辞書エントリーを抽出する (FC)
- 出現頻度分布の比較で不適切エントリーとして抽出された辞書エントリーに対して言い換え表現の存在推定を行う。言い換え表現が存在した場合リスクスコアを推定し、不適切エントリーとしての妥当性を再評価する (WP1)
- 出現頻度分布の比較で適切エントリーとして抽出された辞書エントリーに対して言い換え表現の存在推定を行う。言い換え表現が存在しなかった場合リスクスコアを推定し、適切エントリーとしての妥当性を再評価する (WP2)

これらの手順をもとに作成した実験システムの概要を図 1 に示す。

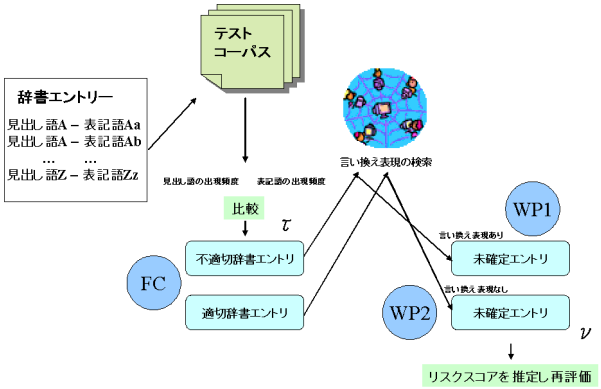

図 1: 実験システムの概要

テストコーパスには MEDLINE データ中の 1996 年以降の論文抄録 3755182 文書を用いた。また、言い換え表現の存在推定のための Web 文書検索には Google Web API[1] を用いた。

5 実験結果と考察

前章で述べたデータ、システムを用いて不適切辞書エントリーの抽出実験を行い、人手正解判定および表記語の出現頻度で定義されるリスクスコアと照らし合わせ、不適切辞書エントリーの抽出評価を行った。各手法による不適切辞書エントリー抽出の精度、再現率、F 値は表 3 のようになった。

表 3: 各手法による結果

	精度	再現率	F 値
FC	0.816	0.587	0.683
FC+WP1	0.821	0.587	0.685
FC+WP2	0.786	0.656	0.715
FC+WP1+WP2	0.791	0.656	0.717

出現頻度分布の比較 (FC) に加え言い換え表現の存在推定を考慮 (WP1、WP2) することによって F 値が改善していることがわかる。不適切辞書エントリーの再評価 (WP1) においては再現率が下がっていないことから、適切な辞書エントリーのみを再評価で抽出できたことがわかる。不適切辞書エントリー抽出ではできるだけ多くの不適切エントリーを抽出することが重要となるため、再現率と F 値の両方を

[1] http://www.google.com/apis/

改善できていることから本研究で提案した方法は有用であることがわかる。

人手評価による不適切辞書エントリーの中には、Web 検索において得られた文書集合中に言い換え表現を持つものがある。このため、出現頻度分布の比較で不適切と判定された辞書エントリーの再評価で適切エントリーを抽出するために言い換えパターンを拡張すると前述のような不適切エントリーも抽出してしまい、精度が改善されても再現率を下げる可能性がある。したがって、言い換え表現があるにも関わらず専門家が不適切と判定した辞書エントリーについてははなぜ違いが生じたのかを吟味する必要があると思われる。

また、出現頻度分布の比較で不適切と判定された辞書エントリーに対する再評価では、言い換え表現が存在しないことを推定しているが、Web 検索で得られた文書サンプルからの推定では不十分であると考えられる。また、ハイフン挿入の有無や大文字小文字の違う (例えば見出し語”Pit1”,”LEP”に対する表記語”Pit-1”,”Lep”) といった文字列が類似している辞書エントリーについては適切辞書エントリーであっても言い換え表現が記述される可能性は少ない。そのため適切エントリーに対する言い換え表現を用いた再評価 (WP2) では再現率は改善されるが、本来適切であるエントリーも不適切としてしまい精度が下がっている。言い換え表現の抽出のみで精度を落とさずに大幅に再現率を上げることは難しく、文字列を比較し類似しているエントリーは別途評価することなどが必要であると考えられる。

6 まとめ

本稿ではまず、ライフサイエンス向けテキストマイニングの自然言語処理で用いられる、複数の外部リソースから作成する専門用語辞書中に含まれる見出し語・表記語のペアからなる不適切エントリーの問題点について述べた。その上で、見出し語および表記語の出現頻度分布の比較、言い換え表現の存在推定に基づいて不適切辞書エントリーを抽出する手法を提案し、実験によりその有効性を検証した。実験の結果、出現頻度分布の比較および言い換え表現の存在推定を行うことによって再現率と F 値の両方を改善でき、できるだけ多くの不適切エントリーを抽出することが求められる本研究が対象としている目的において、提案手法は有用であることがわかった。しかしながら精度を大幅に落とさず、さらに再現率を上げていくにはさらなる改善手法が必要であることもわかった。

謝辞

本研究で用いた実験用データの作成にあたって協力していただいた東北化学薬品株式会社 生命システム情報研究所 峯岸大輔氏、北嶋洋志氏に感謝いたします。また本研究にあたって多大なるサポートをいただいた東北化学薬品株式会社 生命システム情報研究所 小岩弘之所長、日本 IBM 株式会社 東京基礎研究所 武田浩一氏に感謝いたします。

参考文献

[1] B.Boeckmann, A.Bairoch, R.Apweiler, M.C.Blatter, A.Estreicher, E.Gasteiger, et al. SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, 31(1), 365–370, 2003.

[2] B.L.Humphrey and H.M.Schoolman. The Unified Medical Language System: An Informatics Research Collaboration. *Journal of the American Medical Informatics Association*, 5(1), 1–11, 1998.

[3] A. Koike and T. Takagi. Gene/ Protein/ Family Name Recognition in Biomedical Literature. *HLT-NAACL 2004 Workshop: BioLink 2004, Linking Biological Literature, Ontologies and Databases*, 9–16, 2004.

[4] M. Krauthammer and G. Nenadic. Term Identification in the Biomedical Literature. *Journal of Biomedical Informatics*, 37(6), 512–526, 2004.

[5] T. Nasukawa and T. Nagano. Text analysis

and knowledge mining system. *IBM System Journal*, 40(4), 967-984, 2001.

[6] K.D.Pruitt and D.R.Maglott. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research*, 29(1), 137–140, 2001.

[7] Y. Tsuruoka and J. Tsujii. Probabilistic Term Variant Generator for Biomedical Terms. *Proceeding of the SIGIR 2003*, 167–173, 2003.

[8] H. Shatkay and R. Feldman. Mining the Biomedical Literature in the Genomic Era: An Overview. *Journal of Computational Biology*, 10(6), 821–855, 2003.

[9] O. Tuason, L. Chen, H. Liu, J.A.Blake and C. Friedman. Biological nomenclatures: a source of lexical knowledge and ambiguity. *Proceeding of the Pac. Symp. Biocomput.*, 238–249, 2004.

[10] N. Uramoto, H. Matsuzawa, T. Nagano, A. Murakami, H. Takeuchi and K. Takeda. A Text-Mining System for Knowledge Discovery from Biomedical Documents. *IBM System Journal*, 43(3), 516-533, 2004.