# CRF를 이용한 생물/의학 전문용어 인식

배영준[O1] 김재훈[2] 옥철영[1] 최윤수
[1]울산대학교, [2]한국해양대학교, [3]한국과학기술정보원
young4862@ulsan.ac.kr, jhoon@hhu.ac.kr, okcy@ulsan.ac.kr, armian@kisti.re.kr

# Biomedical Terminology Recognition using CRF

Young-Jun Bae[O1], Jae-Hoon Kim[2], Cheol-Young Ock[1], Yun-Soo Choi[3]
[1]University of Ulsan, [2]Korea Maritime University, [3]Korea Institute of Science and Technology Information

**요 약**

전문용어의 수가 급증하면서 전문용어를 자동으로 인식하는 연구가 활발히 진행되고 있다. 전문용어를 인식하기 위해서 전문용어의 범위를 정한 뒤 그 전문용어의 분야를 선택해야 한다. 본 논문에서는 생물/의학 사전정보와 CRF(Conditional Random Fields) 기계학습 기법을 사용하여 연구를 진행한다. 기계학습을 위한 자질로 품사, 접사, 대소문자, 숫자, 특수문자, 단서어휘 등을 사용한다. 특히 단서어휘와 사전정보를 중요한 요소로 생각하여, 3가지 방법으로 나누어 실험한다. 총 분야의 개수는 7개이며, 각 분야별로 정확률, 재현율, F-measure를 측정한다. 경계인식은 83.92%의 정확률, 96.42%의 재현율, 89.73의 F-measure가 결과로 나타났고, 분야분류는 79.29%의 정확률, 91.06%의 재현율, 84.77%의 F-measure가 결과로 나타났다.

주제어: 전문용어 인식, CRF, 생물/의학 전문용어

## 1. 서 론

전문용어는 전문적 개념을 지칭하는 어휘 또는 어휘의 집합을 말한다. 이러한 전문용어는 각 분야의 전문문서에 사용되지만, 문서 내의 단어 또는 어휘가 전문용어인지 선별해 내는 것은 그리 쉬운 일이 아니다[1]. 지식정보가 기하급수적으로 증가함에 따라, 전문분야 문서와 전문분야 개념과 이를 지칭하는 전문용어도 기하급수적으로 증가하고 있다. 이러한 전문용어 증가양상 때문에 구성원간의 의사소통의 장애, 컴퓨터간의 정확한 정보전달 장애 등 많은 어려움이 발생되고 있다. 이를 해결하기 위해 전문용어 추출 및 인식에 관한 연구가 활발히 이루어지고 있다. 전문용어 인식은 전문용어 후보가 될 용어의 경계를 인식한 후, 문맥 또는 분야에 적당한 전문용어를 선별해 내는 것을 뜻한다. 전문용어 선별 방법으로 사전을 기본적으로 이용하지만, 이 방법으로는 새롭게 만들어지는 전문용어의 파악에는 한계가 있다. 그래서 최근에는 주로 용어의 빈도수 등과 같은 통계적 방법, 언어의 형태, 어휘, 문맥 등의 자질을 기반으로 한 기계학습 방법, 사람의 직관과 여러 가지 패턴을 이용한 규칙적 방법을 사용한다.

본 논문에서는 생물/의학 분야의 전문용어 자질을 바탕으로 Conditional Random Fields(CRF)기법을 이용한 전문용어 인식 방법을 제안한다.

## 2. 관련 연구

전문용어 자동 인식 연구는 크게 규칙 기반 연구와 통계 기반 연구, 그리고 앞의 두 연구방법을 병행하는 혼합형 연구로 구분할 수 있다. 규칙 기반 연구는 사전이나 규칙을 사용하는 방법으로 사람의 수작업을 통한 규칙의 정확률과 사전의 크기가 인식의 정확률을 결정짓는 요소이다. 통계 기반 연구는 지도식 학습과 비지도식 학습으로 나눠진다. 지도식 학습은 사람의 판단을 통해 만들어진 대량의 말뭉치가 준비되어 있을 때 사용하기 좋은 방법이고, 비지도식 학습은 소량의 말뭉치를 대상으로 초기 규칙을 학습·인식의 과정을 반복해 성능을 향상시키는 방법이다. 일반적으로 비지도식 학습보다 지도식 학습이 좋은 인식 결과를 보인다. 학습에 사용되는 통계 모델은, 은닉 마코프 모델(hidden markov model), 신경망(neural network), SVM(support vector machine), 최대 엔트로피 모델(maximum entropy model)등이 있다.

국내외의 최근 전문용어 인식 경향은 다음과 같다. 특정 분야 하나에 대해서 약어, 이형태, 접사 등을 세밀하게 분석한다[2, 3, 4]. 그리고 형태적, 구문적, 의미적 중의성을 발생시키는 용어에 대해서 집중적으로 처리한다[5]. 혹은 2개 혹은 여러 개의 언어를 대상으로 전문용어를 분석해 정리하는 연구가 발표되었고[6, 7], 최근에는 전문용어 인식의 기반이 되는 사전을 구축하는 연구도 진행되고 있다[8].

현재 국외에서는 생물/의학분야에서 개체명과 전문용어를 혼용해서 사용하고 있다. 일반적으로 개체명은 PLO(Person, Location, Organization)을 나타내지만, 국외에서는 세포명, 유전자명, 질병명 등도 개체명으로 보는 경향이 있기 때문에 개체명과 전문용어가 혼용되고 있다.

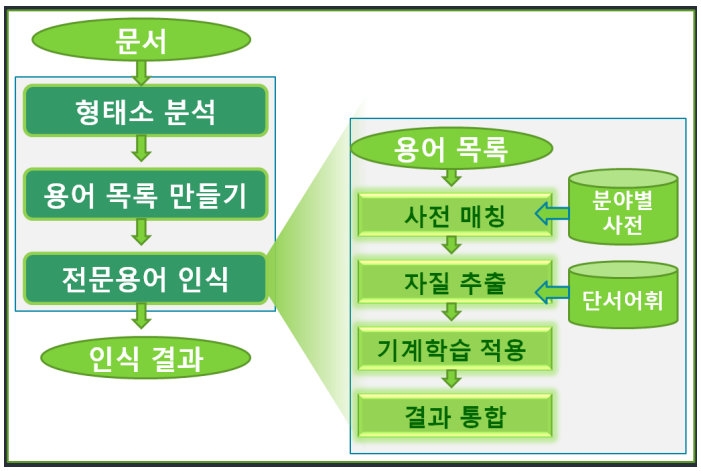
## 3. 전문용어 인식

전문용어 인식을 위해 경계인식과 분야분류 두 단계가 필요하다. 두 단계를 나누어서 각각 기계학습을 실행할 수도 있다. 하지만 본 논문에서는 분야정보를 세부적으로 분류하지 않았기 때문에 분야수가 많지 않다. 그래서 경계인식과 분야분류 두 단계를 하나로 묶어서 기계학습을 수행하였다. 경계인식은 B, I, O 태그를 사용한다. 전문용어가 단일어이면 B는 전문용어의 시작과 끝을 나타내며, 복합어이면 B는 시작을 나타내며, I는 중간과 끝을 나타낸다. O는 일반어를 나타낸다. 전문용어의 인식결과 태그는 다음과 같이 '(B, I)-분야영어명'으로 표기하였다. 생물/의학분야의 분야 종류는 유전자, 단백질, 질병, 세포, 조직 등 몇 가지로 나눌 수 있다. 본 논문에서는 KISTI에서 제공한 생물/의학분야 문서 내의 전문용어 분야 개수 총 7가지를 사용한다. 분야정보는 <표 1>과 같다.

<표 1> 생물/의학 분야 7가지

| 번호 | 영어명 | 한글명 |
|---|---|---|
| 1 | Disease | 질병 |
| 2 | Drug | 약 |
| 3 | Gene | 유전자 |
| 4 | Organism | 생물·조직 |
| 5 | Protein | 단백질 |
| 6 | TechTerm | 기술용어 |
| 7 | Others | 기타 |

(그림 1)은 본 논문에서 제안한 전문용어 인식을 위한 시스템의 구조도이며 구체적인 설명은 이하의 절에서 기술할 것이다.




(그림 1) 전문용어 인식 시스템 구조도

### 3.1 자질

기계학습을 하기 위해 품사(f1), 전체가 대문자인지 여부(f2), 대소문자가 섞여있는지 여부(f3), 숫자 포함 여부(f4), 특수기호 포함 여부(f5), 접두사 2개(f6), 접미사 3개(f7), 전문용어 복합어 내의 앞에 나오는 F단서어휘 여부(f8), 전문용어 복합어 내의 중간에 나오는 M단서어휘 여부(f9), 전문용어 복합어 내의 마지막에 나오는 E단서어휘 클래스(f10), 총 10가지 자질을 사용하였다.

<표 2> 자질의 종류와 표현 형식

| 자질 | 수 | 표현 형식 |
|---|---|---|
| f1 | 36 | NN, NNS, NP, NPS, JJ, ... |
| f2 | 2 | +, - |
| f3 | 2 | +, - |
| f4 | 2 | +, - |
| f5 | 2 | +, - |
| f6 | - | Di, Al, Ch, pn, sy, an ... |
| f7 | - | dia, ity, ion, mia, ncy, ... |
| f8 | 2 | +, - |
| f9 | 2 | +, - |
| f10 | 7 | Disease, Drug, Gene, Organism ... |

생물/의학분야 전문용어 중 소문자로만 구성된 용어가 절반 이상이지만, 특정 질병, 단백질, 약, 유전자의 이름에 해당하는 전문용어는 대문자, 숫자, 특수기호 등을 포함한 경우가 많다. (예:5HT2a, BHC110/LSD1, CaMKII) 그래서 대소문자가 섞여있는지 여부를 묻는 자질과 숫자, 특수기호의 포함 여부를 묻는 자질을 사용하였다. 생물/의학분야의 전문용어는 알파벳 20자 이상의 용어들이 상당히 많기 때문에 약어로 표현될 경우가 많기 때문에 모두 대문자인지 여부를 묻는 자질을 사용하였다. 또한 생물/의학분야의 전문용어는 그리스 라틴어원의 용어들이 많다. 이런 용어들은 접두사, 접미사와 같은 형태소로 구성되며, 특별한 의미를 가진다. 아래의 <표 3>는 특정 접두사, 접미사가 가지는 의미와 예를 보여준다.

<표 3> Disease, Drug 관련 접두사, 접미사

| 접사 | 의미 | 예 |
|---|---|---|
| -algia | pain | talalgia *ankle* |
| -cele | hernia | gastro*cele stomach* |
| -dynia | pain, swelling | uro*dynia urine* |
| -gen | producing, beginning | carcino*gen cancer* |
| -oma | tumour | adeno*ma gland* |
| -osis | abnormal condition | dermatosis skin |
| sulfa- | antibiotic | sulfacetamide |
| cef-<br>ceph- | cephalosporin antibiotic | cefaclor; cephalexin |

접사의 길이는 유동적이다. 그래서 다음과 같이

◈ ('-iasis', '-osis') => '-sis'

◈ ('-dynia', '-penia') => '-nia'

특징만 뽑아낼 수 있도록 접미사는 3개의 알파벳, 접두사는 2개의 알파벳을 자질로 사용하였다. 전문용어 중 복합어 형태가 있다. 이러한 전문용어들을 각 분야의 전문용어 사전마다 어절별로 나누어서 제일 앞에 나오는 어절(F단서어휘), 중간에 나오는 어절(M단서어휘), 마지막에 나오는 어절(E단서어휘)로 나누어 빈도별로 정리한 것이 단서어휘이다. 아래의 (그림 2)는 TechTerm에 대한 단서어휘이다.

TechTerm_word : 테이블

| ID | WORD | F | M | E | OWN | SUM |
|---|---|---|---|---|---|---|
| 132 | therapy | 0 | 1 | 63 | 2 | 66 |
| 145 | treatment | 1 | 2 | 22 | 2 | 27 |
| 7 | test | 0 | 1 | 16 | 1 | 18 |
| 31 | technology | 0 | 0 | 16 | 1 | 17 |
| 155 | surgery | 2 | 1 | 13 | 2 | 18 |
| 45 | therapies | 0 | 0 | 12 | 2 | 14 |
| 73 | technique | 0 | 0 | 11 | 1 | 12 |
| 62 | response | 1 | 1 | 9 | 1 | 12 |
| 21 | scans | 0 | 0 | 8 | 1 | 9 |
| 153 | growth | 2 | 1 | 7 | 1 | 11 |
| 239 | migration | 0 | 0 | 7 | 1 | 8 |
| 202 | transplant | 0 | 1 | 7 | 2 | 10 |
| 401 | transplantation | 0 | 0 | 7 | 2 | 9 |
| 162 | transplants | 0 | 0 | 7 | 1 | 8 |
| 80 | responses | 0 | 0 | 6 | 1 | 7 |
| 2 | chemotherapy | 0 | 1 | 6 | 2 | 9 |
| 88 | mutation | 0 | 0 | 6 | 2 | 8 |

(그림 2) TechTerm에 대한 단서어휘

단서어휘는 본 논문의 시스템에 자질로 중요한 역할을 한다. 특히 복합어인 전문용어 중 마지막 단어는 경계인식 뿐 아니라, 분야 결정에 중요한 역할을 한다. 그래서 그 단어가 단서어휘 사전에 포함되어 있는지에 대한 여부뿐 아니라, 분야정보도 자질 값에 포함시켰다.

단서어휘의 형태는 대소문자, 숫자, 특수기호, 약어, 복수형태 등의 유무와 차이로 인해 여러 가지 이형태들이 나타난다.

예) (Hodgkin Disease), (HODGKIN DISEASE),
(Hodgkin's Disease), (Hodgkin's disease),
(Disease, Hodgkin)

이러한 요소는 시스템의 재현율을 떨어뜨리는 원인이 될 수 있다. 그래서 본 논문에서는 전체가 대문자, 숫자, 기호인 단어를 제외하고 나머지 단어에 대해서 일반화 작업을 하였다. 위에서 제외한 단어는 약어로 볼 수 있는 단어이며, 일반화 작업으로 인해 다른 단어와 중의성이 발생할 수 있기 때문에 제외하였다. 일반화 작업은 우선 단어를 모두 소문자로 교체하고 복수형인 단어를 단수형으로 교체한다.

### 3.2 CRFs(Conditional Random Fields)

CRFs는 조건부 확률을 최대화하기 위해 훈련된, 방향성이 없는 그래프 모델이다. 선형 체인 CRF 모델은 유한 상태 기계에 대응되는 그래프 구조를 가지며, 연속 레이블링에 적당하다. 매개변수를 갖는 선형 체인 CRFs는 아래의 식과 같이 입력 순열 x가 주어진 상태 순열 y에 대한 조건부 확률로 정의된다.

$$P_\lambda(y|x) = \frac{1}{Z(x)} exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(y_{t-1}, y_t, x, t)\right) \quad (1)$$

(식 1)에서 y는 출력상태 (예를 들어, B-Gene, I-Disease 등)를 나타내며, x는 앞에서 선정한 자질들을 나타낸다.

## 4. 실험 및 평가

본 논문에서는 CRF++ 도구를 이용하여 학습과 실험을 하였다. 이 도구는 C++로 구현되어 있으며 LBFGS 알고리즘을 이용하여 빠르게 수렴하도록 학습을 수행한다. 실험은 전문용어의 경계인식과 분야분류를 동시에 수행하였고, 인식결과를 분석하여 경계인식과 분류에 대한 각각의 정확률, 재현율, F-measure를 측정하였다. 정확률은 경계인식 및 분류된 용어의 개수에서 올바르게 인식 및 분류된 용어의 비율을 나타내고, 재현율은 정답문서에 나타나는 각 분야별 모든 전문용어 중 인식된 전문용어의 비율을 나타낸다. F-measure는 정확률과 재현율을 통합적으로 나타내는 평가 기준으로 사용하였다.

$$\text{정확률} = \frac{\text{정확하게 분류된 용어의 개수}}{\text{시스템에서 분류된 용어의 개수}}$$

$$\text{재현율} = \frac{\text{정확하게 분류된 전문용어의 개수}}{\text{정답문서에 나타난 전문용어의 개수}}$$

$$F-measure = \frac{2 \times \text{정확률} \times \text{재현율}}{\text{정확률} + \text{재현율}}$$

### 4.1 실험데이터

KISTI에서 제공한 생물/의학 분야의 논문과 저널 354개의 문서(8,303문장)를 실험데이터로 이용하였다. 이 문서는 전문용어와 개체명에 대해 총 10가지 분야로 수작업 태깅된 문서이다. 이 중 개체명분야(Person, Location, Organization)를 제외한 나머지분야(Disease, Drug, Organism, Protein, TechTerm, Others)를 대상으로 문서를 9:1의 비율로 나누어 학습 및 실험 데이터로 사용하였다. 사전으로 BioLexicon과 GeneOntology 명사 사전을 사용하였다.

### 4.2 실험결과 및 평가

<표 4> 전문용어 경계인식 결과

| Bound | 정확률 | | | 재현율 | | | F-measure | | |
|---|---|---|---|---|---|---|---|---|---|
| | 사전: X 단일: X | 사전: X 단일: O | 사전: O 단일: O | 사전: X 단일: X | 사전: X 단일: O | 사전: O 단일: O | 사전: X 단일: X | 사전: X 단일: O | 사전: O 단일: O |
| B, I | 87.23 | 75.94 | 83.92 | 67.61 | 84.59 | 96.42 | 76.17 | 80.03 | 89.73 |

<표 5> 전문분야별 전문용어 분야분류 결과

| Class | 정확률 | | | 재현율 | | | F-measure | | |
|---|---|---|---|---|---|---|---|---|---|
| | 사전: X 단일: X | 사전: X 단일: O | 사전: O 단일: O | 사전: X 단일: X | 사전: X 단일: O | 사전: O 단일: O | 사전: X 단일: X | 사전: X 단일: O | 사전: O 단일: O |
| Disease | 87.78 | 82.67 | 82.49 | 74.19 | 79.50 | 95.87 | 80.42 | 81.05 | 88.68 |
| Drug | 60.53 | 51.22 | 90.48 | 25.00 | 34.24 | 92.94 | 35.36 | 41.04 | 91.69 |
| Gene | 84.62 | 39.24 | 71.88 | 22.92 | 64.58 | 95.83 | 36.07 | 48.82 | 82.14 |
| Organism | 84.13 | 83.59 | 73.91 | 74.65 | 77.70 | 95.07 | 79.11 | 80.54 | 83.16 |
| Protein | 64.71 | 47.85 | 86.89 | 49.62 | 75.19 | 79.70 | 69.35 | 58.48 | 83.14 |
| TechTerm | 69.35 | 57.91 | 82.06 | 65.71 | 87.14 | 87.14 | 67.48 | 69.58 | 84.53 |
| Others | 64.73 | 55.94 | 75.69 | 46.50 | 70.19 | 86.29 | 54.12 | 62.26 | 80.64 |
| 전 체 | 76.47 | 65.65 | 79.29 | 59.32 | 73.11 | 91.06 | 66.81 | 69.18 | 84.77 |

실험은 다음의 2가지 경우를 고려하였다.

㉠ 사전을 이용한 경우

㉡ 단서어휘를 일반화한 경우

위 2가지 경우를 바탕으로 3가지의 결과를 정리하였다.

(1) ㉠과 ㉡ 모두 사용하지 않은 실험

(2) ㉡만 사용한 실험

(3) ㉠과 ㉡ 모두 사용한 실험

전문용어 경계인식의 결과는 <표 4>와 같이 나타났다. 정확률은 (1)의 경우가, 재현율은 (3)의 경우가, F-measure는 (3)의 경우가 가장 좋은 결과를 보였다. (1)의 경우 87.23%로 정확률은 높았지만 재현율이 현저히 낮은 것을 볼 수 있다. 사전을 사용하지 않았기 때문에 실험데이터 내에 많은 전문용어를 찾지 못했지만, 찾은 것에 대해서는 대부분 정확히 경계를 인식했다고 볼 수 있다. (2)의 경우 정확률은 (1)의 경우보다 약 11%가량 떨어졌지만 재현율은 약 19%가량 오른 것을 볼 수 있다. 단서어휘의 일반화가 재현율에 많은 영향을 주는 것을 알 수 있다. (3)의 경우 사전과 단서어휘 일반화를 동시에 사용할 때 전반적으로 높은 성능을 보였다.

전문용어 분야인식의 결과는 <표 5>와 같이 나타났다. 각 분야별로 정확률, 재현율, F-measure는 정리하였다. 분야분류에서는 모두 (3)의 경우가 좋은 성능을 보였다. 전반적인 양상은 경계인식의 결과와 비슷했다 그러나 결과에서 볼 수 있듯이 분야별로 정확률과 재현율 차이가 현격히 차이가 나는 것을 볼 수 있다. 그 이유는 실험 말뭉치 내에 각 분야별로 어휘수가 많이 차이가 났기 때문으로 볼 수 있다.

지도학습을 위해서는 정확하게 만들어진 대용량의 태깅된 말뭉치가 있을수록 좋은 성능을 낸다. 하지만 본 논문에서는 학습말뭉치가 약 7,000문장에 불과해 학습으로 인한 많은 성능 향상을 바라기는 힘들었다. 그래서 단서어휘 일반화 및 사전에 재현율이 보다 많은 영향을 받은 것 같다.

일반적으로 생물/의학분야 전문용어 인식 연구에 Genia 말뭉치가 많이 사용된다. Genia 말뭉치는 생물/의학과련 2,000개의 논문 요약(abstract)으로 이루어져 각 분야별로 태깅되어있다. 분야명이나 종류가 조금 상이하지만 본 연구에서 수정된 Genia 말뭉치를 추가해 사용했다면, 보다 보편적이고 좋은 성능을 낼 수 있었으리라고 본다. 그리고 Genia 말뭉치를 사용한 다른 논문의 결과들과 비교평가 할 수 있을 것이다.

## 5. 결론

본 논문에서는 CRF 기계학습기법과 사전을 이용한 생물/의학 분야 전문용어 인식 방법을 제안하였다. 사전의 이용 여부, 단서어휘의 일반화 여부에 따라 다른 결과가 나왔다. 단서어휘의 일반화를 통해 정확률은 10% 떨어졌지만 상대적으로 재현율이 13% 올라 F-measure가 약 3% 올랐다. 그리고 단서어휘의 일반화와 전문용어 사전을 이용했을 때 F-measure가 84.77%로 최고의 성능을 보였다.

향후 전문용어 인식의 성능향상을 위하여 보다 많은 전문용어가 태깅된 말뭉치를 이용할 것이다. 특히 분야

범위와 세부적인 내용은 다르지만, Genia Corpus 3.02를 사용하여 학습을 위한 말뭉치를 확장해 보다 보편화된 성능을 끌어낼 것이다. 그리고 사전을 기반으로 한 이형태들의 약어 및 변이형태를 세부적 규칙으로 설정하고 시스템에 적용시킬 것이다.

## 감사의 글

본 연구는 한국과학기술정보연구원에서 수행하는 교육과학기술부 차세대 정보유통 핵심기술 연구개발 사업의 위탁연구로 수행되었습니다.

## 참고문헌

[1] 국립국어원, 전문용어 연구, 태학사

[2] X. Wang, Rule-based protein term identification with help from automatic species tagging, Proceedings of CICLING, pp. 288-298, 2007.

[3] A. Naoakiou, A term recognition approach to acronym recognition, Proceedings of the COLING/ACL 2006, pp. 643-650, 2006.

[4] H. Liu, and C. Blouin and V. Keselj, An unsupervised method for extracting domain-specific affixes in biological literature, Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, pp. 33-40, 2007.

[5] X. Wang and M. Matthews, Species disambiguation for biomedical term identification, Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, pp. 71-79, 2008.

[6] E. Lefever, L. Macken and V. Hoste, Language-independent bilingual terminology extraction from a multilingual parallel corpus, Proceedings of the 12th Conference of the European Chapter of the ACL, pp. 496-504, 2009.

[7] D. Bernhard, Multilingual term extraction from domain-specific corpora using morphological structure, Proceedings of the 9th Conference of the European Chapter of the ACL, pp. 171-175, 2006.

[8] Y. Sasaki, S. Montemagni, P. Pezik, D. Rebholz-Schuhmann, J. McNaught, and S. Ananiadou, BioLexicon: A lexical resource for the biology domain, Proceedings of the Third International Symposium on Semantic Mining in Biomedicine, 2008.

[9] G. Nenadic, I. Spasic and S. Ananiadou, Automatic acronym acquisition and management with domain specific texts, Proceedings of the LREC-3, pp. 2155-2162, 2002.

[10] G. Nenadic, I. Spasic and S. Ananiadou, Terminology-driven mining of biomedical literature, Journal of Bioinformatics, Vol. 19, No. 8, pp. 938-943, 2003.

[11] S. Ananiadou and G. Nenadic, Automatic terminology management in biomedicine, Text Mining for Biology and Biomedicine, pp. 67- 97, 2006.