SVM을 이용한 중국어 고유명사 식별에서의 자질 선택

김풍[1] 나승훈[2] 강인수[3] 리금희[4] 김동일[5] 이종혁[6]
포항공대 정보통신대학원 정보처리학과[1] 포항공대 컴퓨터공학과[2,3,4,6]
중국연변과학기술대학 언어공학연구소[5]
{maple, nsh1979, dbaisk, ljj, jhlee}@postech.ac.kr[1,2,3,4,6], dongil@postech.ac.kr[5]

Feature Selection for Chinese Named Entity Recognition using SVM

Feng Jin[1], Seung-Hoon Na[2], In-Su Kang[3], Jin-Ji Li[4], Dong-Il Kim[5], Jong-Hyeok Lee[6]
Dept. of Graduate School for Information Technology, POSTECH1, Dept. of Computer Science & Engineering, POSTECH[2,3,4,6], Language Engineering Institute, YUST[5] China

요 약

"고유명사 식별"은 사전에 등록되어 있지 않은 고유명사를 찾아내고 분류하는 과정으로 주로 인명, 지명, 조직명을 처리 대상으로 한다. 처리할 데이터는 점점 많아지고 고유명사는 수시로 생겨나기 때문에 고유명사 식별은 정보검색, 질의응답, 기계번역시스템의 핵심 기술 중의 하나로 부각되었다. 고유명사 식별에 있어 정확률과 더불어 식별속도와 식별모듈의 크기가 시스템의 성능에 미치는 문제도 쟁점이 되고 있다. 본 논문에서는 SVM과 자질선택을 결합한 다양한 실험을 통하여 중국어 고유명사의 식별 효율을 높이는 방법을 연구하였다.

1. 서 론

근래에 들어서 사람들은 정보가 넘친다는 표현을 자주 사용한다. 그 이유는 정보의 양의 증가로 인한 것 외에 새로운 용어가 예전에 비해 빈번하게 생성되기 때문이다. 수많은 고유명사를 일일이 사전에 기입하는 것은 현실성이 없으므로 고유명사의 식별은 정보검색, 질의응답 및 기계번역에 있어 필수적인 기능으로 인정되었다. 특히 중국어 문장에는 단어와 단어의 영역을 알려주는 기호가 존재하지 않고, 대부분 중국어 문자가 단어로도 사용이 가능하기 때문에 영어나 한국어에 비하여 중국어 고유명사 식별에는 특히 어려움이 많다.

고유명사 식별에 관하여 많은 연구가 진행되어 왔지만 식별률과 식별속도를 동시에 만족시키지는 못하고 있다. 특히 각광을 받고 있는 SVM(support vector machine)을 고유명사 식별에 접목했을 때 식별률은 만족스러우나 식별속도가 느리고 식별모듈이 크기 때문에 실용성이 높지 않은 상황이다. 그 이유는 고유명사 식별에 있어 어휘정보를 비롯하여 너무 많은 양의 정보가 자질로 사용되었기 때문이다.

보다 실용성 있는 고유명사 식별모듈을 얻기 위해 자질을 선택적으로 사용하는 것이 필수적이다. 본 논문에서는 자질 선택과 SVM의 식별률, 속도, 모듈 크기 사이의 관계를 알아보았다.

본문의 순서는 다음과 같다. 두번째 장에서는 고유명사 식별에 관한 기존 연구들을 설명하였고, 세번째장에서는 자질선택의 방법론과 본 연구에서의 적용방법을 살펴보겠다. 네번째 장에서는 고유명사 식별의 전반적인 과정에 대한 설명이 있고, 다섯번째 장에서는 실험을 통하여 고유명사 식별에 있어 자질선택의 의미를 알아보도록 한다. 마지막 장에서는 본 연구에 대한 결론과 향후 연구 방향에 대한 설명이 있다.

2. 기존연구

개체명 인식에 관련된 연구에서 규칙(Petasis, 2001)과 Decision Tree(Sekine,1998), Maximum Entropy(Hai Leong, 2003), Hidden Markov Model(GuoDong, 2002), Support Vector Machine(Takeuchi, 2002), 그리고 Unsupervised Learning(Collins, 1999) 등의 방법들이 사용되었는데, 그 중에서 근래에 들어서 SVM을 이용한 방법론들이 돋보이는 성능을 나타내고 있다.

(Takeuchi 2002)에서는 분자생물학 영역에서 인명, 단백질명, 화학공식, 컴퓨터코드 등을 식별하는 연구를 하였는데 사용이 가능한 자질의 개수에 대한 제한이 있는 HMM에 비하여 다양한 자질을 사용했을 경우 SVM은 4% 이상의 성능우위를 보였다. 그리고 "polynominal kernel"을 2로 하고, 윈도우 크기는 3으로 하였을 때 가

장 좋은 결과를 얻을 수 있음을 실험을 통하여 보여주었다.

(Juhong, 2004)에서는 단어발음변환 시스템에서 SVM을 중국어 미등록어 식별에 접목하였는데, 人名과지명의 식별성능이 각각 89.49%, 91.14% (F-measure)로 기존 시스템에 비하여 좋은 결과를 얻었다.

고유명사 식별에 있어 다른 많은 통계적 기법에 비하여 SVM의우월성은 검증되었지만, 사용하기에는수행속도가 너무 느리다는 것이 밝혀졌는데 (Hideki 2002)에서는 비슷한 사양의 컴퓨터에서 실행되었을 때 전통적인 규칙기반 시스템은SVM을 사용한 시스템보다 수십배 빨랐다. 이 문제의 해결 방안으로 (Hideki 2002)는 SVM의 "quadratic kernel"식을 단순화 시킨 XQK (eXpand the Quadratic Kernel)을 개발하였고 SVM-Light3.50 보다 무려 102배 빠른 수행속도를 보였다.

SVM Light[1]를 사용하여 고유명사를 식별할 때는 각 고유명사들(인명, 지명, 조직명)의 시작 단어(혹은 문자), 중간 단어(혹은 문자), 마지막 단어(혹은 문자)를 각각 인식하는 모듈이 필요한데 9개 모듈과 전처리로서 고유명사 문자 탐지과정에서 필요한 한 개의 모듈을 추가하면 SVM모듈이 총 10개이다. SVM Light를 사용하고 어휘 10만개, 품사 44개, 의미코드 3,300개 등 자질을 선택하며 윈도우를 3으로 설정한 상태에서 주변 정보의 출현 순서를 고려하여 100만 예제를 학습하는 실험을 하였는데, 얻어진 10개 모델의 크기는 총 92.1MB였다. 이 모델을 학습하여 얻어내는데 많은 시간이 필요할 뿐더러 사용되었을 때 시스템에 상당한 부하를 주고 있음을 발견하였다.

성능을 일정하게 유지하는 전제 하에서 차원을 축소하여 시스템의 부하를 줄이는 것이 본 연구의 주된 목적이다.

3. 차원 축소

고유명사식별과정에서 자주 사용되는 자질 중에 어휘, 의미코드 등은 수량이 상당히 많으므로 학습, 분류에서 예제들의 복잡성으로 인하여 분류의 성능을 상당히 저하시킨다. 분류(Classification)문제에서 차원을 축소하기 위해 자주 사용되는 값은 Document Frequency (DF), Information Gain(IG), Mutual Information(MI), ChiSquare-statistic(CHI), Term Strength(TS) 등이 있는데 (Yiming, 1997)에서는 IG와 CHI가 가장 좋은 결과를 보여주었다. 먼저 IG와 CHI에 관하여 알아보겠다.

1) Light version 5.0 http : //svmlight. joachims.org

Information Gain (IG)

IG는 어떤 자질을 선택했을 때 감소되는 엔트로피의 양을 기준으로 자질들의 질 혹은 유용성을 측정하는 방법으로 기계학습 분야에서 자주 사용된다. 식(3-1)에서 C는 카테고리를 가리키고 m는 카테고리의 개수를 의미한다. 분류문제에서 카테고리의 수는 작은 경우 두개이고 수만개인 경우도 있다. 특정 자질의 질은 해당 자질의 각 카테고리에서 얻은 값은 평균치를 이용하여 판단한다.

(3-1)

$$\begin{aligned} G(t) &= \text{Entropy}(S) - \text{Expected Entropy}(S_t) \\ &= \{-\sum_{i=1}^{m} P(c_i)\log P(c_i)\} - \\ &\quad [P(t)\{-\sum_{i=1}^{m} P(c_i \mid t)\log P(c_i \mid t)\} + \\ &\quad P(\bar{t})\{-\sum_{i=1}^{m} P(c_i \mid \bar{t})\log P(c_i \mid \bar{t})\}] \end{aligned}$$

ChiSquare-statistic(CHI)

CHI는 자질, 카테고리 사이의 의존도를 의미하는데 서로 독립인 카테고리와 자질의 CHI 값은 0이다. 식(3-2)는 CHI를 구하는 식인데, t는 자질을 표시하고 c는 카테고리를 의미한다. A는 t와 c가 공동 출현한 회수이고, B는 c가 아닌 다른 카테고리에서 t가 출현한 회수이다. C는 t를 제외하고 c가 출현한 회수이고, D는 C와 D가 동시에 출현하지 않는 회수이다.

$$\chi^2(t,c) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (3\text{-}2)$$

자질의 CHI는 한 자질이 각 카테고리와의 CHI를 합하여 구하거나(식 3-3), 가장 높은 결과가 나온 CHI를 최종 결과로 선택하는 방법(식 3-4)이 있다.

$$\chi^2_{AVG}(t) = \sum_{i=1}^{m} P(c_i)\chi^2(t,c_i) \quad (3\text{-}3)$$

$$\chi^2_{MAX}(t) = \max_{i=1}^{m}\{\chi^2(t,c_i)\} \quad (3\text{-}4)$$

자질선택의 효과를 확인하기 위해 본 연구에서는 WEKA1-SVM을 사용하였다. SVM은 기타 분류 기법들에 비하여 입력 자질이 많을 경우에도 안정적인 성능을 나타내고 보편적으로 적용할 수 있는 모델을 개발할 수 있다는 장점이 있다. 구분하기 어려운 데이터를 잘 분류하고 학습속도는 비록 느리지만 학습된 모델의 크

기가 작고 수행 속도가 매우 빠르다. SVM은 이진 분류기이므로 문제가 멀티 클래스일 경우에는 여러 개의 이진 분류기들을 모아서 멀티 클래스 분류기를 만들어 해결한다.

사전에 수록되지 않은 단어들은 세그먼테이션 단계에서 문자 단위로 분리되기 때문에 세그먼테이션을 마친 후 대부분의 고유명사는 문자 토큰들로 남게 된다. 고유명사가 앞뒤 단어들과 겹치는 경우도 있는데 (김풍, 2004)에서는 전처리로서 고유명사와 겹칠 확률이 높은 단어를 탐지하고 처리하는 방법을 제시하였다. 본 연구에서는 모든 고유명사들이 중국어 문자 단위로 잘라졌다는 전제 하에 실험이 진행된다. 보통 중국어 문장 세그먼테이션 결과 중에서 문자 단위로 세그먼트가 연달아 있을 경우에 해당 문자들을 대상으로 고유명사 식별이 시작되는데, 기존 중국어 고유명사 식별 논문들에서 사용된 자질은 중국어 문자, 주변 단어, 주변 단어의 품사 등이 일반적이고, 적용 가능성이 있는 자질을 정리하여 보면 다음과 같은 것이 있다.

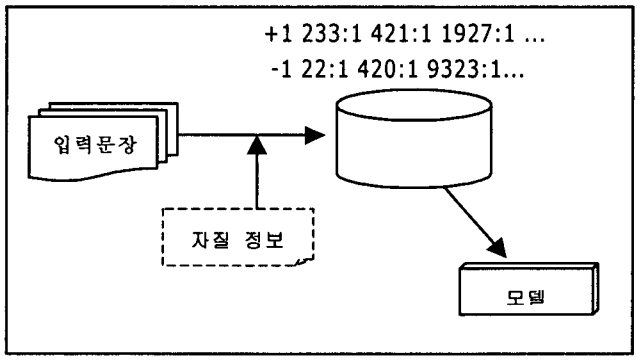
1. 세그먼트된 문자.
2. 세그먼트된 문자의 타입.
   문장기호, 숫자, 일반 한자 등 3가지
3. 세그먼트된 문자의 개수
4. 단어. 세그먼트 문자 앞뒤로 일정한 윈도우 안에 들어간 단어들.
5. 단어의 품사
6. 단어가 속하여 있는 구(phrase)의 문법적 카테고리
7. 단어의 의미부류
   同意詞詞林[2](梅家駒, 1985) 혹은 HOWNET[3]을 참조 할 수 있다.
8. 고유명사 후보가 문장에서의 위치

이상의 자질 중에서 SVM 분류기의 계산 복잡도를 증가시키는 자질들은 1번, 4번, 7번이다. '세그먼트된 문자'는 모든 중국어 한자 문자(6,700개)가 후보이고,"同義詞詞林"의 코드체계는 소분류까지 총 1,428개의 '의미부류'로 이루어진다. 그리고 '단어'의 후보는 사전에 수록된 모든 단어이기 때문에 대략 후보가 몇만 개에서 수십 만개가 있을 수 있다. 그러므로 '자질선택'의 대상인 '단어', '한자 문자', '의미부류' 중에서 중점은 '단어'이다.

4. 고유명사 식별 과정

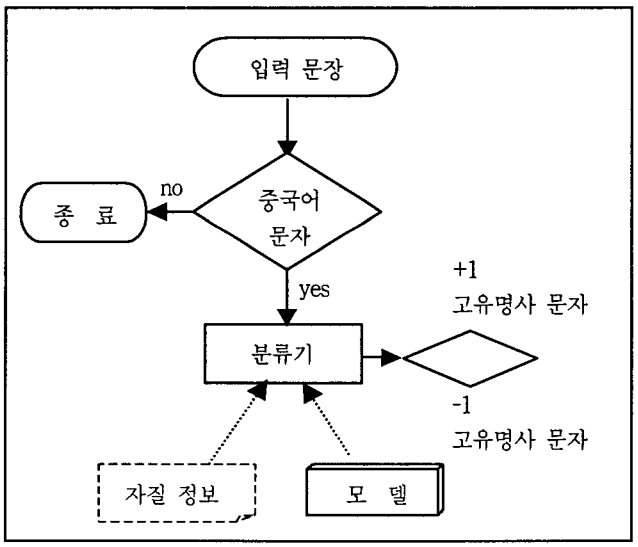
고유명사 식별을 위해서는 준비 단계인 모델 학습 단계(그림 1)와 고유명사 문자 탐지 과정(그림 2) 클래스 지정 과정(그림 3)이 필요하다. 특정 고유명사 후보의 관련 정보가 담겨 있는 벡터를 생성하는 것은 학습단계와 식별단계에서 공통으로 필요하다. 학습단계에서 SVM 학습모듈에 양의 예제와 음의 예제 다수를 입력으로 하여 식별모듈을 얻는데 각 모듈에 들어가는 양의 예제는 타 모듈들의 음의 예제로 사용된다.

기존 연구(Juhong 2004)에서는 고유명사를 이루는 문자들의 위치 클래스를 사용하였는데 [NE-B], [NE-I], [NE-E], [NN] 등(B : 고유명사의 시작 문자, I : 고유명사의 중간 문자, E는 고유명사의 마지막 문자, NE : 인명, 지명, 혹은 조직명, NN : 일반 단어; LMR tag라



[그림 1] SVM model 학습

고도 함)이 있다. 그리고 고유명사 후보 문자에 대한

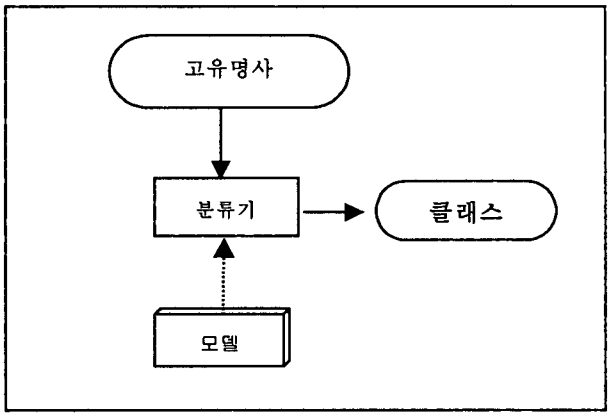


[그림 2] 고유명사 문자 탐지

클래스 부여는 Viterbi Algorithm(Takeuchi 2002) 혹은

2) WEKA version 3.4.2 (University of Waikato 에서 개발)
http : //www.cs.waikato.ac.nz/~ml/weka/
3) HOWNET (중국과학원에서 개발)
http : //www.keenage.com/html/e_index.html

규칙(Juhong 2004)을 사용하였다. 이런 방법을 사용하면 모듈을 생성, 로딩 그리고 Viterbi Algorithm의 많은 계산량으로 인하여 고유명사 식별속도가 상당히 느려질 뿐더러, 하드웨어 공간을 많이 점유하게 된다. 본 연구에서는 고유명사 문자 탐지과정에서 고유명사와 주변 단어 사이의 경계를 확정하고 문자 토큰으로 잘려진 고유명사를 단 한번의 분류를 거쳐 클래스를 지정하였다.



[그림 3] 고유명사 클래스 지정

(그림 3)은 중국어문장 "老○/教/王/小/明/英○"- (선생,가르치다,왕,소,명,영어) 에서 인명 王小明(왕소명)을 식별하여 내는 과정을 보여주고 있다.

```
step 1- 입력문장 : 老○教王小明英○
step 2- 세그먼테이션 & 태깅 :
        T0 (老○) 품사 : n    의미부류 : Ae131
        T1 (教) 품사 : v     의미부류 : Hg051
        T2 (王) 품사 : n     의미부류 : Af051
        T3 (小) 품사 : a     의미부류 : Eb012
        T4 (明) 품사 : a     의미부류 : Eb181
        T5 (英○) 품사 : n    의미부류 : Dk063
step 3- 자질 부여 : (윈도우 크기 = 3)
        T0 1235 : 1 17001 : 1 17981 : 1 ...
        T1 239 : 1 17002 : 1 17873 : 1 ...
        T2 422 : 1 17001 : 1 1836 : 1 ...
        T3 29 : 1 17003 : 1 17922 : 1 ...
        T4 167 : 1 17003 : 1 18720 : 1 ...
        T5 2017 : 1 17001 : 1 19525 : 1 ...
step 4- 고유명사 문자 탐지
        老○ 教/v 王/NE 小/NE 明/NE 英?
step 5- 고유명사 클래스 지정
        老○/n 教/v 王小明/nr 英○/n
```

[그림 4] 고유명사 식별 예제

5. 실험 및 결과 분석

실험 코퍼스로 인민일보(1998년 6개월, 약 50MB; 7,814,225 단어)를 사용하였다. 인민일보는 사람이 수동으로 품사를 태깅한 코퍼스이고 다양한 도메인의 내용들이 내포되어 있다. 코퍼스 중에 40MB를 학습코퍼스 사용하였는데 인명이 35,654개(109,083회), 지명이1,306개(34,376회), 단체명이 440개(4,843회), 그리고 기타 개체명이 4,787개(18,432번)가 나타났다. 학습을 위한 벡터는 클래스마다 약 백만 개가 만들어졌고 각 벡터에 들어 갈 수 있는 자질 공간은 10여 만개이다. 성능평가 단계에서는 나머지 10MB 인민일보 코퍼스를 사용하였다. 시스템사양은 Pentium 2.4GB, RAM 512MB이다.

먼저 고유명사 후보의 추출성능, 자질선택과 효율의 관계를 살펴보겠다. (표1)과 (표2)는 가장 많은 자질을 소유하는 단어 자질들에 CHI와 IG를 각각 적용하였을 때의 성능의 변화 상황을 보여준다. 실험결과에서 F-measure은(식 5-1)에서 얻어질 수 있다.

$$F-measure=\frac{2\times precision\times recall}{precision+recall} \quad (5\text{-}1)$$

CHI	Feature Num.	Model Size (MB)	F-measure (%)	Run Time (sec.)
0.0	110,373	8.77	98.59	5897
0.1	82,779	7.33	98.46	5267
0.2	64,142	6.61	98.37	4812
0.3	41,054	5.72	98.24	4298
0.4	34,815	5.37	98.09	3821
0.5	22,074	4.80	98.01	3354
0.6	19,370	4.42	97.80	2790
0.7	15,724	3.71	97.47	2293
0.8	12,123	3.29	97.26	1828
0.9	10,023	2.88	96.57	1360
1.0	8,599	2.59	96.22	926

[표 1] CHI를 사용했을 경우

IG	FeatureNum	Model Size (MB)	F-measure(%)	Run Time (sec.)
0.0	110,373	8.77	98.58	5644
0.1	78,627	7.28	98.42	5108
0.2	58,045	6.45	98.35	4659
0.3	40,286	5.79	98.11	4153

0.4	26,026	4.93	97.86	3684
0.5	17,258	4.19	97.42	3177
0.6	14,512	3.62	97.21	2730
0.7	12,383	3.48	96.97	2092
0.8	9,134	2.73	96.40	1631
0.9	8,240	2.38	96.17	1117
1.0	7,399	2.24	95.87	712

[표 2] IG를 사용했을 경우

위의 두 표를 대조하면 CHI 값이 IG 보다 전반적으로 좋다는 것을 알 수 있다.

계속되는 실험은 클래스지정 결과이다. 실험에서는 고유명사 후보 추출에서 좋은 성능을 보인 CHI를 사용하였다.

CHI	Feature Num	Model size (MB / %)	F-measure(%)	Run Time (sec / %)
0.0	110,373	6.59 / 100.0	98.14	4573 / 100.0
0.4	26,026	4.35 / 66.01	98.87	1878 / 41.07
0.8	9,134	2.01 / 30.50	97.39	1280 / 27.99
1.0	7,399	1.56 / 23.67	96.41	647 / 14.15

[표 3] 인명 식별 평균 성능

CHI	Feature Num	Model size (MB / %)	F-measure(%)	Run Time (sec / %)
0.0	110,373	6.25 / 100.0	96.38	4346 / 100.0
0.4	26,026	4.16 / 66.56	96.48	1612 / 37.09
0.8	9,134	1.85 / 29.60	95.27	1011 / 23.26
1.0	7,399	1.37 / 21.92	94.86	595 / 13.69

[표 4] 지명 식별 평균 성능

여기서 한가지 더 고려하여 볼 수 있는 자질은 윈도우 밖의 자질의 사용이다. 그 중에서도 기사의 타이틀과 문단의 첫 문장이 기사의 내용 전달에서 가장 중요한 것 만큼 고유명사 식별에서 자질로서 가치가 있을 수 있다. 기사 제목과 문단의 첫 마침표 혹은 쉼표가 출현할 때까지의 내용을 고유명사 식별의 자질로 사용하고 CHI를 1.0으로 정하여 지명 식별실험을 하였는데 앞의 실험에 비하여 오히려 성능이 떨어졌다.

Feature Num	Model size (MB / %)	F-measure (%)	Run Time (sec / %)
23,810	3.57 / 57.12	93.98	1472 / 34.16

[표 5] 지명 문자 식별 평균 성능

5. 결론 및 향후연구

이상으로 SVM을 이용하여 고유명사를 식별하는데 있어 자질 선택의 방법과 효과를 살펴보았다. SVM은 많은 자질을 사용하는 것을 허용하는 장점이 있다. 하지만 많은 자질을 사용함에 따라서 학습속도가 현저하게 느려지고, 모듈의 크기와 수행속도도 자연언어처리 시스템에 상당한 부하를 주게 된다.

본 논문에서는 고유명사 탐지와 클래스 지정 단계에서 자질 선택 방법으로서 CHI와 IG을 적용하였는데 CHI를 사용했을 때가 IG를 사용했을 때보다 좋은 성능을 보였고, CHI threshold를 0.8로 정하였을 때 고유명사 탐지 성능이 98% 이상 인명 식별성능이 97% 이상 지명 식별 성능이 95% 이상 도달하였다. 모델의 크기와 수행시간도 모든 자질을 사용했을 경우의 23.26% ~ 29.60%로 줄어들은 것을 발견할 수 있다.

마지막 실험에서는 윈도우 밖의 정보를 클래스 지정 단계에 사용하였는데 사용하기 전에 비하여 모델이 많이 커졌지만 성능은 오히려 조금 떨어진 것을 발견할 수 있다. 이는 고유명사와 멀리 떨어져있는 정보는 고유명사 식별의 자질로 사용하기에는 적절하지 않다는 것을 설명한다.

본 논문에서 제안한 고유명사 식별방식은 LMR 태그 방식을 사용하지 않기 때문에 수행시간은 상당히 단축되었으나, 식별성능은 초기 고유명사 문자 탐지 결과의 정확성에 많이 의지하게 된다. 향후 고유명사 문자의 탐지 성능을 높이는 것을 목적으로 더 많은 연구를 진행할 계획이고 다른 통계적 방법과 자질선택 방법을 결합하는 실험을 통하여 실용성이 높은 중국어 고유명사 식별 방법을 연구하고자 한다.

6. 참고 문헌

G. Petasis, F. Vichot, F. Wolinski, G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, "Using Machine Learning to Maintain Rule-based Named - Entity Recognition and Classification Systems". In Proceedings of the 39th Conference of Association for Computational Linguistics (ACL-EACL 2001), pp. 418 - 425, July 9 - 11 2001, Toulouse, France

S. Sekine and R. Grishman and H. Shinnou, A Decision Tree Method for Finding and Classifying Names in Japanese Texts, In Proceedings of the Sixth Workshop on Very Large Corpora, 1998

C. Hai Leong, & N. Hwee Tou (2003). Named Entity

Recognition with a Maximum Entropy Approach. Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003). (Shared Task Paper). (pp. 160-163). Edmonton, Alberta, Canada.

Z. GuoDong and S. Jain, Named Entity Recognition Using a HMM-based Chunk Tagger, ACL2002. Philadelphia . July 2002

K. Takeuchi and N. Collier. Use of support vector machines in extended named entity recognition. In Proceedings of the sixth Conference on Natural Language Learning (CoNLL-2002)

M. Collins and Y. Singer (1999) : Unsupervised models for named entity classification, in Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora

J. Ha, Yu Zheng, Gary Geunbae Lee, "High Speed Unknown Word Prediction Using Support Vector Machine For Chinese Text-to-Speech Systems", IJCNLP 2004

Hideki Isozaki, "Efficient Support Vector Classifiers for Named Entity Recognition", COLING 2002

Y. Yang, "A Comparative Study on Feature Selection in Text Categorization", Proceedings of the Fourteenth International Conference on Machine Learning, 1997

김 풍, "SVM을 이용한 중국어 개체명 식별", 한국정보과학회 2004 봄 학술발표논문집(B), 제31권 제1호, pp.934-936

梅家駒, "同義飼飼林", 上海辭書出版社, 1985