SVM을 이용한 중국어 개체명 식별

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

Recognition Of Chinese Named-Entity Using Support Vector Machine

Feng Jin[1]. Seung-Hoon Na[2]. In-Su Kang[2]. Jin-Ji Li[2], Dong-Il Kim[3]. Jong-Hyeok Lee[2]
Dept. of Graduate School for Information Technology, POSTECH[1], Dept. of Computer Science &Engineering, POSTECH[2]. Language Engineering Institute, YUST[3] China, and Advanced Information Technology Research Center(AITrc)

요 약

본문에서는 최근 들어 각광을 받고 있는 패턴인식 방법론인 Support Vector Machine을 이용하여 중국어 개체명를 식별하는 방법을 제안하고자 한다. SVM(support vector machine)은 입력 자질이 많을 경우에도 안정적인 성능을 나타내고 보편적으로 적용할 수 있는 모델을 개발할 수 있는 장점이 있다. 실험에서 어휘, 품사, 의미부류 등 많은 수의 자질을 이용하였다. 실험결과는 본문에서 제안한 방법이 튜닝을 거치지 않아도 좋은 성능을 나타낼 수 있고, 수행 속도도 만족스럽다는 것을 보여주었다.

1. 서 론

미등록어는 기계번역, 정보검색 시스템의 성능에 아주 큰 영향을 미친다. 날로 늘어나는 미등록어를 전부 사전에 기입하는 것은 현실적으로 불가능한 일이므로 미등록어의 자동식별은 필수적인 기능중의 하나이다. 미등록어에는 복합어와 개체명이 포함되는데 개체명의 범주에는 인명, 지명, 단체명, 시간, 날짜 등이 있다. 합성어 처리만으로도 복잡하고 어려 가지 문제가 존재하므로 복합어에 관련된 연구는 본문에서 다루지 않고, 개체명 인식에 주안점을 두고자 한다.

개체명 인식에 관련된 기본연구에는 전문가에 의하여 만들어진 규칙을 기반으로 하는 방법 (Petasis, 2001)과 Supervised Learning방법, 예를 들면 Decision Tree(Sekine,1998), Maximum Entropy(Hai Leong, 2003), Hidden Markov Model(GuoDong, 2002), Support Vector Machine(Takeuchi, 2002), 그리고 Unsupervised Learning(Collins, 1999) 이 있다.

이상의 방법론 중에서 SVM이 가장 뛰어난 성능을 보이고 있다(Isozaki, 2002). 본 논문에서는 SVM을 이용하여 중국어 문서 안의 인명, 지명, 단체명 등 중국어 개체명의 인식 정확도를 높이는 방법을 제안하고자 한다.

1 SVM

SVM은 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 크기가 작고 수행 속도가 매우 빠르다. SVM은 이진 분류기이므로 문제가 멀티 클래스일 경우에는 여러개의 이진 분류기들을 모아서 멀티 클래스 분류기를 만들어 해결한다.

복잡도를 줄이고 overtraining을 방지하기 위하여 대부분의 NER(Named Entity Recognizer) 알고리즘은 아주 제한된 자질만을 사용하지만 SVM은 높은 차원의 vector를 쉽게 처리할 수 있다. 우리는 어휘정보, 의미부류 등 자질을 포함한 방대한 자질집합을 이용하는 멀티 클래스 분류기를 구축하여 중국어 개체명 식별을 수행하였다.

2. 학습

2.1 자질(feature)

본 시스템에서 각 학습 패턴은 어휘, 품사, 의미부류 등 자질들로 이루어진 벡터이다. 많은 자질을 처리할 수 있는 SVM의 특성을 이용하여 본 연구에서는 단어들을 그룹지을 수 있는 자질을 가능한 많이 사용하고자 한다. 후보로 선택할 수 있는 자질은 다음과 같은 것들이 있다. (표 1)

1)단어

실험용 사전에 수록된 한개 한자 이상의 한자로 구성된 단어는 155,689개이고 한자는 6,763개이다.

2)품사(단어일 경우)

북경대학에서 정의한 품사 태그 45개를 사용한다.

3)병음

성조를 제외한 한자 병음 집합에는 서로 다른 병음이 모두 408개 있다. ex) ba, bo, bi, ...

4)의미코드

의미코드는 " 同義詞詞林'(梅家駒, 1985)의 코드체계에서 소 분류까지 이용하는데 총 1,428개의 의미부류로 이루어진다.

5)토큰의 길이

문장을 세그먼테이션하고 생성된 각 토큰 안에 들어 있는 한자의 개수이다.

자질	표기
단어	W
품사	P
병음	Y
의미부류	S
토큰 길이	L

표 1 자질의 표기

세그먼테이션을 진행한 다음 사전에 등록되지 않는 개체명은 보통 글자단위로 분리되는데 한 글자짜리 토큰이 연속으로 이어지는 경우가 개체명일 가능성이 가장 크다. 그러므로 연속으로 이어지는 각 한자 토큰이 개체명에 속하는지를 판단하는 것이 가장 중요하다. 예를 들어 연속되는 한 글자짜리 토큰 " A B C" 중에서 A가 인명의 시작 즉성(surname)이고 B와 C는 각각 인명의 중간과 끝으로서 이름(given name)이라는 것을 발견할 수 있다면 인명식별은 성공한 것이다.

2.2 클래스(class)

한자 토큰을 분류하기 위하여 우리는 개체명의 클래스를 13개 정의하였다. 인명에서 시작, 중간, 끝을 각각 RB, RM, RE 3개 클래스로 나누었고 지명, 단체명, 기타 개체명도 각각 같은 방법으로 시작, 중간, 끝 3개 클래스로 나누었다. 그리고 한 글자단위로 된 토큰이지만 개체명에 속하지 않는 것은 OTHER라는 클래스를 부여하였다. (표 2)

클래스	기호		클래스	기호
인명 시작	RB		단체명 시작	TB
인명 중간	RM		단체명 중간	TM
인명 끝	RE		단체명 끝	TE
지명 시작	SB		기타 개체명 시작	ZB
지명 중간	SM		기타 개체명 중간	ZM
지명 끝	SE		기타 개체명 끝	ZE
한글자 단어	OTHER			

표 2 클래스를 표기하는 기호

4. 개체명 식별 알고리즘

개체명 인식과정은 다음과 같다. (그림 1 참조)

(1)세그먼테이션, 태깅 (step 1)

입력문장을 TOTAL-Tagger[1]를 이용하여 문장을 세그먼테이션하는데 왼쪽에서부터 오른쪽으로의 최장일치법을 사용한다. 세그먼테이션된 토큰에 품사, 의미정보, 병음 등 정보를 태깅한다.

(2)개체명 시작 토큰 체크

세그먼테이션 된 문장에서 개체명은 한 글자로 된 토큰으로 이루어졌다는 전제하에서 한 글자로 된 토큰들을 확인한다. 하지만 개체명에 두 글자 이상으로 이루어진 일반 단어가 포함되는 경우가 있다. 예를 들면 " 李建国" 는 인명이고 " 建国" 는 ' 나라를 건설하다' 는 의미를 갖는 동사인 동시에 사람 이름에도 아주 흔히 쓰인다. 이런 경우를 대비하여 우리는 학습 코퍼스 안의 개체명에 내포된 일반 단어를 추출하였고, 개체명 식별 단계에 앞서 그런 단어들이 발견되면 한자 단위로 세그먼트를 시킨다. 그리고 그 토큰들이 개체명의 일부분으로 판단되지 않으면 다시 원래 상태로 복원하여 준다.

(3)자질 부여 과정 (step 2)

현재 가리키고 있는 토큰을 T_0라고 하면 좌우 윈도우 사이즈내의 토큰들(윈도우 사이즈가 3일 경우, $T_{-3}, T_{-2}, T_{-1}, T_0, T_{+1}, T_{+2}, T_{+2}$)에 자질을 부여한다. 부여된 자질에는 바이너리 값을 대응시킨다.

자질 부여는 두 글자 이상으로 이루어진 토큰인 경우에는 W, P, S, L 자질들에 값을 부여하고, 한 글자 토큰인 경우에는 W, P, L 자질들에 값을 부여한다. 이것은 한 글자짜리 토큰에 태깅된 품사와 의미코드가 틀렸을 가능성이 크고, 한자는 여러 가지 발음이 있을 수 있지만 두 글자 이상으로 이루어진 단어에 대응되는 발음은 한 가지 밖에 있을 수 없기 때문이다.

(5)클래스 지정 과정 (step 3)

SVM을 이용하여 현재 토큰(T_0)에 13개 클래스 중의 하나를 지정한다.

(6) 개체명 확정과정 (step 4)

연속되어 있는 토큰들의 클래스가 토큰들을 하나의 개체명으로 인식하기에 논리적으로 합당할 경우, 해당 토큰들을 하나로 묶어 주고 태깅한다. 예를 들어 인접한 3개의 한자 토큰이 RB, SM, RE으로 분류되었다면 인명의 중간에는 SM이 올수 없으므로 개체명이 식별될 수 없다.

```
문장에서 인명" 王小明" 의 식별 과정:

입력문장: 老师教王小明英语

step 1- 세그먼테이션 & 태깅:
        (老师) 품사:n   의미부류:Ae131    병음:lao3shi1
        (教)   품사:v   의미부류:Hg051    병음:jiao1
        (王)   품사:n   의미부류:Af051    병음:wang2
        (小)   품사:a   의미부류:Eb012    병음:xiao3
        (明)   품사:a   의미부류:Eb181    병음:ming2
        (英语) 품사:n   의미부류:Dk063    병음:ying1yu3
step 2- 자질 부여:(윈도우 사이즈 = 3)
        T-3 W:NULL  P:NULL          S:NULL     L:0
        T-2 W:老师          P:n     S:Ae131    L:2
        T-1 W:教  Y:jiao1   L:1
        T0  W:王  Y:wang2   L:1
        T+1 W:小  Y:xiao3   L:1
        T+2 W:明  Y:ming2   L:1
        T+3 W:英语          P:n     S:Dk063    L:2
step 3- 클래스 지정 과정
        老师 教/OTHER 王/RB 小/RM 明/RE 英语
step 4- 개체명 확정 과정
        老师 教/OTHER 王小明/nr 英语
```

그림 1 개체명 인식 과정

[1] " 포항공대 지식 및 언어공학 연구실" 의 중국어 태거이다.

5. 실험

본 실험에서 SVM은 잘 알려져 있는 SVM^{Light}[1]를 사용하였다. 학습 코퍼스로 품사 태깅된 인민일보(1998년도, 약 55.5mb; 7,814,225 단어)코퍼스 중에 50mb를 사용하였는데 인명이 35,654개(109,083회), 지명이 1,306개(34,376회), 단체명이 440개(4,843회), 그리고 기타 개체명이 4,787개(18,432번) 나타났다. 학습을 위한 벡터는 클래스마다 약 백만개 만들어졌다. 각 벡터에 들어 갈 수 있는 자질은 총 114만 여개이고, 13개 모델을 학습시키는데 사용한 시간은 총 17,524초(모델 평균 1,348초)이고 생성된 모델의 평균 크기는 5.5MB이다. 시스템 사양은 Pentium 2.4GB, RAM 256MB이다.

실험과정에서는 나머지 0.5mb 인민일보 코퍼스와 인터넷에서 임의로 추출한 뉴스(0.2mb)를 각각 closed domain과 open domain 실험 코퍼스로 사용하였다. 12만여개의 벡터를 분류하는데 사용한 시간은 65초이다. 표3은 실험 결과이다.

	closed (%)			open (%)		
	P	R	F	P	R	F
인명	93.19	82.61	87.58	89.42	80.03	84.46
지명	94.69	90.64	95.62	91.67	90.11	90.88
단체명	88.98	86.08	87.51	87.01	78.04	82.28
기타 개체명	83.49	69.14	75.64	89.40	61.43	72.82
평균	90.09	82.12	85.84	89.38	77.40	82.61

P=precision, R=recall, F=F-score, F=P*R*2/(P+F)

윈도우 사이즈: 3

표 3 개체명 식별 성능

4. 결론 및 향후 연구

SVM을 이용하여 개체명을 중점으로 미등록어를 식별하는 방법을 살펴보았다. SVM은 많은 자질을 사용하는 것을 허용하고 언어에 따라서 특별한 튜닝을 요구하지 않는 장점이 있다. SVM에는 조절 가능한 다양한 커널 변수가 존재하지만 본 실험에서는 SVM^{Light}의 기본값인 값만 사용하였다. 학습데이터에서 인명에는 아세아 인명과 서양인명이 모두 포함되어 있지만 중국어 인명표기 습관상 서양 인명 집합과 아세아 인명을 표시하는 한자 집합은 많이 다르다. 이것은 인명식별의 성능에 부정적인 영향을 끼쳤다. 동양인명과 서양인명을 다른 부류로 정의하고 시스템을 구축하면 인명식별에서 뚜렷한 향상이 있을 것이라고 생각한다. 그리고 " 기타 개체명" 의 recall이 기타 미등록어에 비하여 많이 낮은 것은 다양한 하위 부류가 포함되어 있기 때문이다.

문장 속에서 클래스들의 논리적인 일관성은 클래스기반으로 하는 미등록어 식별에서 관건적인 단계이다(그림 1, step 3,4). 향후 연구에서 이 문제는 Viterbi Search를 이용하여 풀고자 한다. 그리고 더욱 효과적인 자질과 자질의 조합을 연구하고, 기타 통계적인 방법(예를 들어 HMM)과 결합하여 보다 좋은 성능을 갖는 방법론을 개발하고 세그먼테이션, 품사태깅에 적용할 계획이다.

[1] SVM^{Light}는 http://svmlight.joachims.org 에서 구할 수 있다.

6. 참고문헌

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

H. Isozaki, H. Kazawa: Efficient Support Vector Classifiers for Named Entity Recognition Proceedings of COLING-2002, pp.390--396, 2002.

梅家駒, "同義飼飼林", 上海辭書出版社, 1985