기계학습과 사전을 이용한 개체명 세분화

이기중°, 이도길, 임해창, 임수종*

고려대학교 컴퓨터학과, 한국전자통신연구원 음성/언어정보연구센터*
{kjlee°, dglee, rim}@nlp.korea.ac.kr, isj@etri.re.kr*

Fine Grained Classification of Named Entities Using Machine Learning and Dictionary

Kijoong Lee°, Dogil Lee, Haechang Rim, Soojong Lim*

Dept. of Computer Science & Engineering, Korea University,
Speech/Language Technology Research Center of Electronics and Telecommunications Research Institute*

요 약

개체명 인식은 효과적인 정보추출 시스템을 구축하기 위해 반드시 선행되어야 하는 작업이다. 지금까지의 개체명 인식에 관한 연구는 인명이나 조직, 장소와 같은 일반적인 개체명 인식 작업이 대부분이었다. 그러나, 효과적인 정보추출을 위해서는 이런 일반적인 개체명들을 더욱 세분화할 필요가 있다. 본 논문에서는 SVM기반 기계학습법과 기구축된 사전과의 편집거리 비교법을 이용하여 개체명을 세분화하는 방법을 제시한다. 실험은 개체명과 세분화된 범주가 부착된 공연 관련 문서 100개 중 80개는 학습집합, 20개는 실험집합으로 사용하였고 성능 평가 척도는 정확도(accuracy)를 이용해 개별적으로 평가하였다. 실험 결과 기계학습법과 사전을 이용한 방법을 결합한 모델이 가장 좋은 성능(정확도 72.91%)을 보였다.

1. 서 론

전형적인 정보추출 시스템들은 자연어로 된 문서를 분석하여, 사용자가 원하는 정보를 선별하고, 그 결과를 정제되고 가공된 형태로 제시해 준다[1]. 정보추출의 단계는 개체명(Named Entity) 인식 단계, 상호참조(Coreference) 인식 단계, 정보추출 단계로 나눌 수 있다[2]. 지금까지 인명이나 공연장소, 조직 등과 같은 일반적인 개체명(이하 NE) 인식에 관한 연구는 많이 있었다[3]. 그러나 사용자가 원하는 정보를 선별하기 위해서는 NE 인식만으로는 부족한 경우가 많다. 예를 들어, 어떤 사용자가 뉴스기사로부터 공연 정보를 얻고자 한다면, 공연자나 공연단체, 혹은 공연장소에 관련된 정보를 요구할 것이다. 따라서, 인명, 조직, 장소와 같은 NE를 공연자, 공연단체, 공연장소 등으로 더 세분화할 필요가 있다.

본 논문에서는 지지벡터기계(이하 SVM)기반 기계학습법과 사전 항목과의 편집거리 비교법을 사용하여 인명, 조직, 장소와 같은 NE를 공연자, 공연단체, 공연장소와 같은 세분화된 범주로 분류하는 방법을 제안하고자 한다.

2. 관련 연구

NE 인식에 관한 연구는 많이 있었지만, 상대적으로 개체명을 세분화하려는 시도는 많이 없었다. Fleischman은 단어 빈도와 의미 정보 등을 자질로 사용하고, 결정트리나 SVM 등의 기계학습 알고리즘을 이용해 인명(PERSON)을 8개의 세분화된 범주로 분류하고자 하였으며, 시스템의 정확도는 70.4%였다[4].

본 논문에서는 인명 외에도 장소, 조직, 날짜, 시간, 제목과 같은 NE들을 추가적으로 고려하였으며, 기계학습법뿐만 아니라 기구축된 사전을 활용하는 방법도 함께 모색하였다.

3. 개체명 세분화 방법

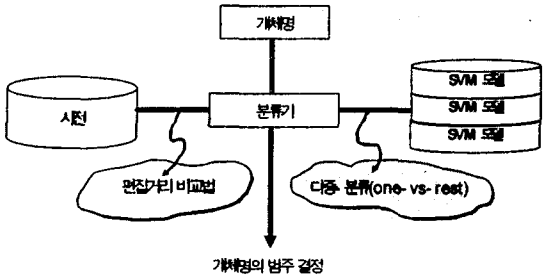
3.1 SVM기반 기계학습법을 이용한 방법

SVM(Support Vector Machine)은 구조적인 위험치 최소화 원리(structural risk minimization principle)에 따라 학습을 수행하는 방법으로 이진분류나 회귀추정에 사용되는 기계학습법이다[5].

SVM기반 기계학습법을 이용한 개체명 세분화 방법은 크게 '학습'과 '분류' 과정으로 나눌 수 있다. 학습 과정에서는 NE의 세분화된 범주가 부착된 학습문서로부터 SVM 학습을 통해 각 범주별로 SVM 모델을 생성한다. 분류 과정은 NE만 부착된 문서를 입력으로 받아 학습을 통해 생성한 SVM 모델에 의해 각각의 NE에 세분화된 범주를 부착한 문서를 생성하게 된다. 실험에 사용한 자질은 다음과 같다.

- 어휘 문맥: 해당 NE의 형태소
- 주변 문맥: 해당 NE의 좌(우)측 n번째에 위치한 형태소
- 문장 문맥: 해당 NE가 나타난 문장에 있는 모든 명사/동사

주변 문맥에서, n은 해당 NE로부터 좌측 또는 우측의 상대적인 위치로서 본 논문에서는 3을 사용하였다. 즉, 해당 NE의 좌우측 세 번째까지의 형태소를 주변 문맥으로 정의했다. 이렇게 추출된 문맥 자질은 벡터로 표현된다. SVM은 기본적으로 이진



[그림 1] SVM기반 모델과 사전기반 모델의 결합

분류만이 가능하기 때문에 본 연구에서는 다중분류가 가능하도록 하기 위해 '하나-대-나머지(one-vs-rest)' 방식을 사용하였다[6]. 하나-대-나머지 방식은 n개의 범주에 대해 n개의 이진 SVM 분류기를 만들고, 각각의 이진 SVM 분류기의 결과값을 이용해 입력 데이터가 어떤 범주에 속하는지 아니면 그 범주를 제외한 나머지 범주에 속하는지를 결정하는 방식이다. 결과값은 0을 기준으로 양수와 음수 값을 가질 수 있다. 이 값이 양수이면서 절대값이 크다면 해당 범주에 속하는 정도가 크고 반대로 결과값이 음수이면서 절대값이 크다면 해당 범주에 속하지 않을 정도가 크다고 말할 수 있다. 하나-대-나머지 방식은 n개의 범주 중에서 각 범주에 대응되는 SVM 분류기의 결과값이 최대인 것을 골라 그 분류기에 대응되는 범주를 선택한다.

3.2 개체명 사전과의 편집거리 비교법을 이용한 방법

기구축된 사전을 이용해 NE를 세분화할 수도 있다. 예를 들어, 장소에 해당하는 NE를 국가, 도시 또는 공연장소로 분류하고자 할 때, 각 범주에 해당하는 사전이 있는 경우 이를 이용해 분류할 수 있다. 일반적으로 NE는 새롭게 만들어지는 특성을 갖고 있기 때문에 사전에 등재되어 있지 않은 경우가 많다. 따라서, 분류하고자 하는 NE와 각 범주에 해당하는 사전 항목들과의 엄격한 매칭을 시도하기보다는 편집거리를 통해 사전 항목과의 유사도를 측정하여 이를 분류에 이용하는 방법이 더 유리하다. 편집거리(edit distance)란 주어진 두 문자열 패턴의 유사도를 비교하는데 사용되는 측정치로서 한 문자열을 다른 문자열로 바꾸기 위해서 필요한 최소한의 문자 삽입, 삭제, 대체 회수로 정의된다[7]. 즉, 편집거리가 작은 값을 가질수록 두 문자열은 유사한 문자열임을 의미한다.

이 방법의 장점은 첫째, 띄어쓰기 오류나 철자오류 혹은 음차과정에서의 불일치문제를 완화시킨다. 예를 들어 "베토벤"이 사전에서 원작자로 분류되어 있을 경우 엄격한 매칭을 수행하면 "배에토밴", "배토밴" 등의 새로 입력된 개체명들과 매칭이 일어나지 않아 분류에 실패한다. 또한, "예술의 전당"이 사전에 공연장소로 등재되어 있을 경우에도 "예술의전당"과 같이 띄어쓰기가 다소 다른 경우 매칭이 일어나지 않게 된다. 하지만 편집거리를 사용하는 경우 이 예들은 작은 편집거리를 갖기 때문에 같은 개체명으로 인식될 확률이 높아진다. 둘째, 공통 접미사를 가진 개체명을 분류하는데 유리하다. 가령, '세종문화회관'이 공연장소 사전에 등재되어 있을 때, '춘천문화회관'은 '문화회관'이라는 공통접미사가 있기 때문에 공연장소로 분류될 확률이 높아진다.

3.3 기계학습법과 편집거리 비교법을 결합한 방법

[그림 1]과 같이 SVM이 분류한 결과와 사전과의 편집거리를 이용해 생성된 결과를 결합하고자 한다. 이를 위해 두 방법으로부터 생성된 결과값을 0과 1사이의 점수값으로 정규화한 뒤, 이 점수를 최종 선택에 반영한다.

SVM으로부터 생성된 결과의 신뢰도 점수를 0과 1사이의 값으로 정규화하기 위해 다음과 같은 수식을 적용한다.

$$\hat{d}_i = \frac{d_i + 1}{2}$$

SVM이 내준 결과값 d_i가 -1보다 작거나 1보다 크면 새로운 변환값 $\hat{d}_i$은 0보다 작거나 1보다 커지게 된다. 그러나 결과값의 최소, 최대값이 정의되어 있지 않기 때문에 정확히 0에서 1사이로 정규화하는 것이 어렵다. 그러나 절대값이 1보다 크다는 것은 그만큼 해당 부류에 속하거나 속하지 않을 정도가 크다는 것을 의미하므로 그 값을 그대로 사용하여 다음과 같은 수식을 적용한다.

$$a = \max_i \hat{d}_i \qquad score_{svm} = a \times \frac{1}{2}\left(1 + \frac{1}{n-1}\sum_{i=1}^{n}(a - \hat{d}_i)^2\right)$$

여기서, a는 변환된 결과값의 최대값이고, n은 범주의 개수이다. SVM 점수는 a를 그대로 사용하는 것이 아니라 다른 범주와의 거리값의 차이를 반영한다. 가령, 거리값이 1, 0.8, 0.7인 경우보다는 1, -0.3, -0.7과 같은 경우에 점수를 더 높게 주기 위한 것이다 단, 최종 점수는 원래 a값의 절반 이하로 감소되지 않도록 하였다. 마지막으로, 점수가 1을 초과하면 1로, 0보다 작으면 0이 되도록 하였다.

사전과의 편집거리를 정규화하여 0과 1사이의 점수값을 만들기 위해 다음과 같은 수식을 사용하였다.

$$k = \frac{1}{\left(\sum_{i=1}^{n}\frac{1}{d_i}\right)} \qquad score_{edit-dist} = \begin{cases} 0 & (d_i \neq 0, kd_i \geq \frac{1}{\theta}) \\ \frac{1}{kd_i} & (d_i \neq 0, kd_i < \frac{1}{\theta}) \\ 1 & (d_i = 0) \end{cases}$$

여기서, d_i는 개체명과 사전 항목과의 편집거리이다. k는 점수를 0과 1사이로 정규화하기 위한 비례 상수이고, n은 범주의 수이다. θ는 점수의 임계값이다. 임계값을 넘지 못하는 점수는 0으로 처리하였고, 편집거리가 0일 경우는 완전히 일치하는 경우이므로 이 때는 점수를 1로 주었다.

최종적으로, 두 개의 결과를 결합하기 위해서 다음과 같은 방법을 사용한다. SVM이 분류한 범주와 편집거리 비교법으로 분류한 범주가 같을 때는 그 범주를 선택한다. 두 개 방법이 분류한 범주가 다를 때는 다음과 같이 각 점수에 가중치를 곱한 후 최종 점수가 높은 쪽의 범주를 선택한다.

$$\alpha \times score_{svm} \ , \ (1-\alpha) \times score_{edit-dist} \quad (0 \leq \alpha \leq 1)$$

여기서, α는 SVM 점수에 대한 가중치이며, 이 값은 0과 1사이의 값이어야 한다.

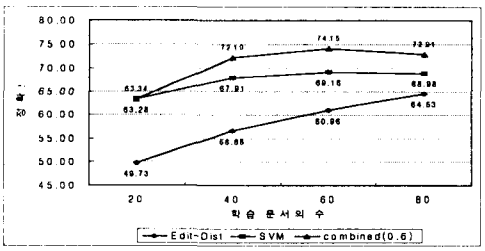
4. 실험 및 평가

실험을 위해 개체명과 [표 1]과 같이 정의한 세분화된 범주를 부착한 공연 분야 신문기사 100문서를 사용하였다.

개체명	세분화된 개체명
PERSON	예술가, 원작자, 기타
LOCATION	국가, 도시명, 공연장소, 기타
ORGANIZATION	공연단체, 기타
TITLE	공연제목, 기타
DATE	연대, 날짜, 기타
TIME	시간, 기타

[표 1] 공연 분야의 세분화된 개체명 범주 정의

첫 번째 실험에서는 SVM기반 모델과 사전기반 모델 그리고 결합모델간의 성능을 비교하였다. 성능 평가의 척도는 정확도(accuracy), 즉, 정답을 얼마나 많이 맞추었는가로 평가하였다. 학습문서의 수에 따라 분류 정확도를 측정한 결과는 [그림 2]와 같다. [그림 2]에서 볼 수 있듯이, 결합 모델이 일관되게 가장 좋은 성능을 보였고, SVM기반 모델이 사전기반 모델보다 성능이 좋았다. 실험에서 사용한 α값은 0.6, θ값은 0.49이다.



[그림 2] 학습문서의 수에 따른 각 모델의 성능비교

두 번째 실험에서는 각 모델의 NE별 정확도를 비교해 각 방법의 기여 정도를 알아보았다. [표 2]에서 볼 수 있듯이, 장소와 조직의 경우 사전기반 모델의 성능이 SVM 기반 모델과 비슷하거나 더 좋은 성능을 보였고, 나머지 NE에 대해서는 SVM 기반 모델의 성능이 대체로 더 좋았다.

NE	범주	학습 문서	편집거리		SVM		Combined(0.6)	
		내의 개수	맞춘개수	정확도	맞춘개수	정확도	맞춘개수	정확도
인명	예술가	103	30	42.04	95	69.43	94	70.70
	원작자	13	0		4		4	
	기타	41	36		10		13	
장소	국가	27	23	62.50	25	62.50	25	70.19
	도시명	28	13		9		13	
	공연장소	35	15		29		28	
	기타	14	14		2		7	
조직	공연단체	46	40	77.92	0	40.26	16	58.44
	기타	31	20		31		29	
제목	공연제목	55	35	59.00	47	67.00	45	65.00
	기타	45	24		20		20	
날짜	연대	22	20	89.13	21	92.39	22	92.39
	날짜	67	60		64		62	
	기타	3	2		0		1	
시간	시간	30	30	96.77	30	96.77	30	96.77
	기타	1	0		0		0	
전체		561	362	64.53	387	68.98	409	72.91

[표 2] 각 모델의 NE별 정확도(학습:80, 실험:20)

결합 모델의 성능이 더 좋아진 이유는 실험 결과에서 보듯이 NE별로 각 모델의 성능이 차이가 나기 때문이다. 예를 들어, 장소와 조직과 같은 개체명은 문맥 정보보다는 개체명 자체의 어휘정보에 더 의존하는데 이 때는 편집거리 비교법이 유리하고, 인명이나 공연제목과 같은 다른 개체명들은 어휘 정보보다는 문맥 정보에 더 의존하므로 SVM기반 기계학습법이 더 유리하다. 결합 모델은 NE에 따라 더 유리한 모델을 선택할 수 있기 때문에 전체적으로 성능이 향상된 것이다.

5. 결론 및 향후 연구

본 논문에서는 SVM기반 기계학습법과 사전과의 편집거리 비교법을 이용해 NE를 세분화된 범주로 분류하는 모델들을 제시하였다. 각 NE별 모델의 성능 평가 실험에서는 기계학습법을 이용한 모델이 사전을 이용하는 방법보다 대체로 성능이 좋았으나, 장소와 조직과 같은 개체명을 분류할 때는 사전을 이용한 모델의 성능이 더 좋았다. 그리고 두 방법의 장점을 취할 수 있는 결합 모델은 학습문서의 수를 달리하였을 경우에도 두 모델보다 일관되게 좋은 성능을 보였다.

향후 연구로는 편집거리 점수를 계산하는 수식에서 길이가 짧은 개체명이 상대적으로 길이가 긴 개체명보다 더 높은 점수를 받지 않도록 개체명 길이로 정규화하여 수식을 수정해야 한다. 또한 SVM 학습에서는 기본적인 문맥 자질만을 사용했으나 빈도나 의미 정보와 같은 추가 자질들을 도입할 수 있을 것이다. 마지막으로 두 개의 방법을 결합할 때 가중치를 자동으로 결정할 수 있는 방법이 고려되어야 한다.

참고문헌

[1] Ralph Grishman, "Information Extraction:Techniques and Challenges", In Proceedings of the Seventh Message Understanding Conference(MUC-7), Columbia, MD, April 1998.

[2] MUC-7 (1998), Proceedings of the Seventh Message Understanding Conference (MUC-7), http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ie_task.html

[3] Bikel, D., Schwartz, R., Weischedel, R., "An algorithm that learns what's in a name. Machine Learning: Special Issue on NL Learning", 34, 1-3, 1999

[4] Fleischman, M. and Hovy, E. "Fine Grained Classification of Named Entities", 19th International Conference on Computational Linguistics (COLING). Taipei, Taiwan, 2002.

[5] SVM-light, http://svmlight.joachims.org/

[6] Hsu and C. Lin, "A comparison of methods for multi-class Support Vector Machines", In IEEE Transactions on Neural Networks, 2002.

[7] Levenshtein, V. "Binary codes capable of correcting deletions, insertions and reversals", Soviet Physics-Doklady 10. pages 707-710, 1966