Self-Organizing Map을 이용한 한국어 동사 클러스터링

박성배°, 장병탁, 김영택
서울대학교 컴퓨터공학과

Korean Verb Clustering Using Self-Organizing Maps

Seong-Bae Park, Byoung-Tak Zhang and Yung Taek Kim
Dept. of Computer Engineering, Seoul National University

요 약

본 논문에서는 목적어-동사 관계의 분포에 따라 한국어 동사를 자동적으로 클러스터링하는 방법을 제시한다. SOM(Self-Organizing Map)이 입력 패턴을 분석하고 가시화하는데 뛰어난 성능을 보이므로, 본 논문에서는 클러스터링하는 방법으로 SOM을 채택하였다. 일단 맵(map)이 만들어지고 나면 학습하는 동안 경험하지 못한 동사도 쉽게 적당한 클러스터로 분류될 수 있고 클러스터들 간의 의미 거리도 맵을 이용하여 쉽게 계산할 수 있다. 본 논문에서 제안한 방법을 명사 확률 분포의 상대 엔트로피(relative entropy)에 기반한 클러스터링 방법과 비교해 본 결과, SOM에 의해 만들어진 동사 클러스터가 상대 엔트로피를 이용해서 만들어진 클러스터를 잘 반영한다는 것을 알 수 있었다.

1. 서 론

자연언어 처리의 통계적 방법에서는 개별 단어보다는 단어 클래스에 대한 통계를 수집함으로써 언어 모델을 보완하는 것이 일반적이다. 단어 클래스가 단어의 의미를 자연스럽게 표현하기 때문에, 단어 클래스는 통계적 언어 모델에서 데이터 부족 문제를 해결하고 언어 모델을 일반화 하는 데 도움을 준다. 그러므로, 단어 클래스는 확률 CFG(context-free grammar)로 씌어진 구문 분석 시스템에서 구문 모호성을 해소하는데 특히 도움이 된다

통계적 언어 모델에서 데이터 부족 문제를 해결하기 위해서, 일반적으로 통계 정보를 수집하는 동안 경험하지 못한 단어의 확률을 이미 알고 있는 비슷한 단어의 확률로 근사시킨다 [4]에서는 단어의 유사성을 상호 정보(mutual information)의 평가치인 공기 집수로 측정하였다 하지만, 이 논문에서는 이 유사성을 이용해서 단어 클래스를 만드는 방법에 대해서 기술하지 않았다. 그리고, [10]에서는 한국어 명사구 대등 접속 구조의 파악을 위해 *공기 유사도*라는 단위를 제시하였다. 이 논문에서는 비록 단어 클래스를 만드는 방법이 제시되었지만, 명사의 유사도라는 개념이 빙사가 가질 수 있는 동사들의 중복성만 고려한 것이기 때문에 동사와 명사가 말뭉치에서 얼마니 자주 같이 쓰이는가 하는 실제적인 분포를 무시하였다

본 논문에서는 동사의 목적어로 쓰이는 명사의 분포에 따라 동사를 클러스터링하는 새로운 방법을 제시한다. 이를 위해서 SOM(Self-Organizing Map)이 사용되었는데, 이는 SOM (Self-Organizing Map)이 입력 패턴을 분석하고 가시화하는데 뛰어난 성능을 보이기 때문이다. 일단 맵(map)이 만들이지고 나면 학습하는 동안 경험하지 못한 동사도 쉽게 직당한 클러스터로 분류될 수 있고, 클러스터들 긴의 의미 거리도 SOM의 특성 때문에 쉽게 계산할 수 있다. 본 논문에서 제시한 방법을 상대 엔트로피를 이용해 단이 클래스를 만드는 모델과 비교하여 본 모델의 유효성을 검증하였다

* 본 논문은 1998년 한국진자통신인구원 위탁과제(104)에 의해 지원받았음

2. 클러스터링 모델

2.1 상대 엔트로피

단이의 유사도를 측정하기 위한 많은 단위들이 제안되어 왔지만 [4, 6, 7, 8, 10], 그 중에서 본 논문에서는 상대 엔트로피 (relative entropy) 혹은 쿨백-라이블러 거리 (Kullback-Leibler distance)를 사용하였다. 두 확률 질량 함수 $p(x)$와 $q(x)$ 사이의 상대 엔트로피는 다음과 같이 정의된다.

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

이 정의에서는 $0 \log \frac{0}{p} = 0$과 $p \log \frac{p}{0} = \infty$을 가정한다.

상대 엔트로피 $D(p \parallel q)$는 항상 0보다 크거나 같고 $p = q$일 때만 0이므로, 이 값을 두 확률 분포 사이의 거리로 생각할 수 있다. 그러나, 일반적으로 상대 엔트로피는 교환 법칙이 성립하지 않고 삼각 부등식을 만족하지 않으므로, 본 논문에서는 상대 엔트로피를 확장하여 아래와 같이 정의된 *확장 상대 엔트로피*를 사용하였다.

$$D'(p \parallel q) = \frac{D(p \parallel q) + D(q \parallel p)}{2}$$

동사 집합 $V = \{v_1, v_2, \cdots, v_n\}$과 명사 집합 $N = \{n_1, n_2, \cdots, n_k\}$에 대해서 $p(v_i)$는 아래와 같이 정의된다.

$$p(v_i) = < p(n_1 \mid v_i),\ p(n_2 \mid v_i),\ \cdots,\ p(n_k \mid v_i) >$$

$$P(n_j|v) = \frac{C(n_j, v)}{\sum_{n_j \in N} C(n_j, v)}$$

이기서, $C(n_i, v)$는 말뭉치에서 n_i와 v가 함께 쓰인 횟수이다.

가능한 모든 동사 쌍에 대해서 상대 엔트로피를 계산한 후, 동사들을 그림 1에 있는 알고리듬에 의해서 클러스터링한다. 동사는 그래프에서 노드로서 표현되고, 그래프 긴선의 가중치는 간선에 의해서 연결된 두 동사 사이의 상대 엔트로피 값이다. 상대 엔프로피 값이 가장 작은 두 노드가 전체에서 의미적으로 가장 가까운 동사들이므로 하나로 묶고, 이 과정을 원하는 개수의 클러스터가 생길 때까지 반복한다. 표 1은 실험적으로 24개의 동시

클러스터 번호	동 사
1	완화하다 철폐하다
2	해독하다 암호화하다
3	팔다 판매하다 거듭하다 되풀이하다
4	극대화하다 향상시키다 약화시키다 회복시키다
5	거론하다 언급하다
6	처치하다 물리치다
7	떠넘기다 회피하다
8	머금다 삼키다 글썽이다
9	고덕이다 숙이다 갸우뚱하다

표 1. 상대 엔트로피를 이용하여 24개의 한국어 동사를 9개의 클래스로 클러스터링한 결과

를 클러스터링하여 9개의 클래스를 얻은 결과이다.

2.2 Self-Organizing Map

상대 엔트로피를 이용하여 단어를 클러스터링하는 방법은 모든 가능한 단어 쌍에 대해서 상대 엔트로피를 구한 후 욕심쟁이 알고리듬(Greedy Algorithm)으로 클러스터링하므로 클러스터링하고자 하는 단어의 수가 늘어날수록 매우 복잡해진다

이 문제를 해결하기 위해서 본 논문에서는 SOM (Self-Organizing Map)이 사용되었다 SOM은 입력 패턴에 따라 다양하게 셀이 조율된 평평한 신경망 배열로 볼 수 있다. SOM에서는 학습 예제가 실 벡터 $\mathbf{x}(t)\in R^n$으로 표현되고, t는 예제에 대한 첨자이거나 학습 시간에 대한 변수이다. 처음에 임의로 초기화된 각 셀 i의 가중치 벡터는 다음 식에 의해서 학습된다.

$$\mathbf{w}_i(t+1)=\begin{cases}\mathbf{w}_i(t)+\alpha(t)(\mathbf{x}(t)-\mathbf{w}_i(t)) & \text{if } i\in N_c(t)\\ \mathbf{w}_i(t) & otherwise\end{cases}$$

여기서 $\alpha(t)\in(0,1]$은 학습율이고, Nc는 승자인 c의 맵 상의 이웃을 뜻한다. 각 입력 벡터의 승자는 입력 벡터와 가중치 벡터 사이의 유사한 정도에 의해서 결정된다. 즉, 입력 벡터 $\mathbf{x}$에 대해서 승자 c는 다음 식에 의해서 결정된다.

$$\|\mathbf{x}-\mathbf{w}_c\|=\min_i\{\|\mathbf{x}-\mathbf{w}_c\|\}$$

SOM에서는 승자 주위의 셀들을 갱신하기 위한 *이웃 함수* (neighborhood function)이 사용되며, 이러한 이웃 함수로는 Mexican-hat function, bubble function, Gaussian function 등 여러 가지가 있다. Gaussian function이 구현하기 쉽고 안정적이므로[5], 본 논문에서는 이웃 함수로서 Gaussin function을 사용하였다.

그림 2는 SOM을 이용하여 표 1의 동사들을 클러스터링한 결과이며, 비교적 표 1을 충실히 반영하였음을 알 수 있다.

3. 말뭉치에서 정보 추출

비교적 복잡하지 않으면서 널리 쓰이는 통계적 언어 모델 중 하나로 n-gram 모델을 들 수 있다 이 모델은 선행하는 $n-1$ 개의 단어만이 다음 단어의 확률에 직접적으로 영향을 끼친다는 가정을 하고 있다. 따라서, n-gram 모델은 구현하기 쉽고 강력하지만 몇 가지 문제도 지니고 있다 우선, n-gram 모델에서는 n-gram의 경계를 넘어선 영역에 대한 정보가 전혀 없다. 더 큰 문제는 이 모델이 문법 역할이 아니라 단어의 나열 순서에 의해 정의되었기 때문에 수집된 데이터의 많은 부분이 실제 문맥을 잘 반영하지 못한다는 것이다 예를 들어, "I met him *at*"과 "I

```
Given n: the number of clusters we want
[step 1] Make a fully-connected weighted graph where
         node = word
         weight of the edge = augmented relative entropy
                              between words connected by the edge
[step 2] <i, j>=two nodes with a minimum augmented
         relative entropy
[step 3] k = a new node resulting from merging i and j
[step 4] for all nodes l such that l ≠ i and l ≠ j and l ≠ k
           weight(l, k) = (weight(l, i) + weight(l, j)) / 2
[step 5] Remove node i and j from the graph.
[step 6] if(# of nodes in graph > n) then goto [step 2]
         else print words in each node.
```

그림 1. 확장 상대 엔트로피에 따라 단어를 클러스터링하는 욕심쟁이 알고리듬(Greedy Algorithm)

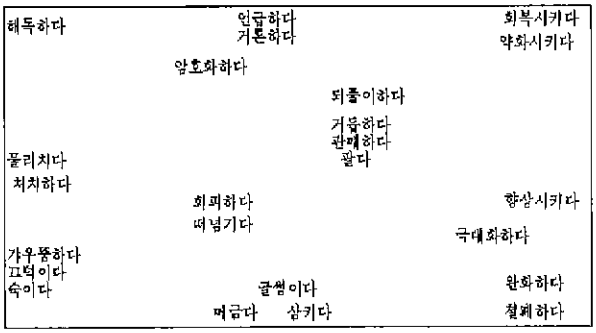

그림 2. SOM으로 표 1의 동사들을 클러스터링한 결과. 이 SOM은 10×10 개의 노드를 가지고 있었다.

met him yesterday *at*"은 전치사 '*at*'을 중심으로 해서 볼 때 비슷한 정보를 가져야 하지만, trigram 모델에서는 전혀 다르게 생각된다.

n-gram 모델 대신으로 본 논문에서는 목적어-동사 관계를 이용하여 동사를 클러스터링하였다. 한국어가 부분 자유 어순의 특성이 있기 때문에, 한국어의 경우에는 이런 모델이 n-gram 모델보다 더 효율적이다. 목적어-동사 관계를 파악하기 위해서는 한국어 파서가 필요하지만,아직 높은 정확도를 갖는 안정적인 한국어 파서가 존재하지 않는다. 그러나, 한국어에서는 조사와 이미로 문법 관계가 표현되며, 특히 목적어 관계는 조사 '을', '를'에 의해서 표현되고 이 조사를 가지는 단어는 가장 가까이 나타나는 동사의 목적어가 되는 경향이 있다. 따라서, 이런 정보를 활용하이 한국이 파서 없이도 목적어-동사 정보를 추출할 수 있다.

4. 실험 및 결과

형태소 분석기와 품사 태거를 이용하여 말뭉치에서 목적어-동사 쌍을 추출하였다. 본 실험에서는 KORDIC(Korean R & D Information Center)에서 개발한 형태소 분석기와 HMM 품사 태거를 사용하였는데, 이들은 품사 태깅에서 최선의 것만 선택해도 98% 이상의 분석 정확도를 보인다[9]. 신문 기사로 이루어진 50만 어절 크기의 말뭉치에서 41,285 개의 목적어-동사 쌍을 추출하였다. 빈도수가 낮은 정보는 통계에서 혼란을 초래하므로,

	Answer should be *yes*	Answer should be *no*
The model says *yes*	*a*	*b*
The model says *no*	*c*	*d*

표 2. 모델 평가를 위한 contingency table

15번 이하의 빈도수를 가지는 동사나 명사를 포함한 쌍을 제거하였다. 이런 쌍들을 제거한 후에, 26,047 개의 쌍이 남았고 이들은 1,531 개의 명사와 987 개의 동사로 이루어져 있었다.

이 목적어-동사 쌍을 학습 예제로 삼아서 위에서 설명한 SOM으로 동사를 클러스터링하였다. 문제의 복잡도를 줄이기 위해서, 1,531 개의 명사를 그림 1에 있는 알고리즘을 이용하여 200 개의 클래스로 분류하였다. 따라서, 987 개의 동사는 각각 200 차원 벡터로 표시되었다. 즉, 입력 벡터 $\mathbf{x}(t) \in R^{200}$는 200 개의 명사 클래스에 대한 조건부 확률 $p(n_i|v)$로 구성된다. SOM은 이 987 개의 동사 벡터에 대해서 학습을 하였다.

Shanon의 정보 이론에 기반한 단어 클러스터링 방법은 단어 클러스터링에 가장 널리 쓰이는 방법들 중의 하나이다. 이 모델에서는 두 동사 v_i와 v_j의 실험적 분포가 비슷하면 비슷한 동사로 간주되며, 비슷하다는 것은 확장 상대 엔트로피 $D'(v_i \| v_j)$로 측정할 수 있다. 확률적인 관점에서 보면, 상대 엔트로피를 이용해서 만들어진 클러스터는 이론적으로 이상적이다. 따라서, 상대 엔트로피를 활용한 모델이 항상 정확한 결과를 보인다는 가정하에, 본 논문에서 제시한 방법을 상대 엔트로피 모델과 비교함으로써 평가할 수 있다 모든 가능한 동사 쌍에 대해서, 두 모델은 각각 두 동사가 같은 클러스터에 속하는지를 결정한다 만약 두 동사가 같은 클러스터에 속하면 각 모델은 '*yes*'라고 대답하고, 아니면 '*no*'라고 대답한다.

본 논문에서 제시한 모델의 성능을 평가하기 위해서 정보 검색이나 심리학에서 널리 쓰이는 contingency table 방법을 이용하였다 이 방법에서는 재현률(recall)과 정확률(precision)이 아래와 같이 정의된다.

$$\text{재현률} = \frac{a}{a+c} \cdot 100\%$$

$$\text{정확률} = \frac{a}{a+b} \cdot 100\%$$

여기서 a, b, c는 표 2에 정의된 것과 같다. 즉, 재현률은 상대 엔트로피 모델이 '*yes*'라고 대답한 것을 바탕으로 할 때 정확한 '*yes*'의 비율을 뜻하고, 정확률은 본 모델이 '*yes*'라고 대답한 것 중 정확한 '*yes*'의 비율을 뜻한다. 재현률과 정확률을 합한 단위인 F-measure는 아래와 같이 정의된다.

$$F = \frac{(\beta^2+1) \cdot \text{재현률} \cdot \text{정확률}}{\beta^2 \cdot \text{재현률} + \text{정확률}}$$

위 식에서 β는 재현률과 정확률 사이의 비중인데, $\beta=1.0$이 본 실험에서 쓰였고 이는 두 단위를 같은 비중으로 본다는 뜻이다.

실험은 다양한 수의 SOM 셀들을 가지고 진행하였다. 재현률과 정확률을 가능한 모든 973,182(= 987×986) 개의 동사 쌍을 대상으로 해서 측정하였으며 그림 3이 그 결과를 보여준다. 재현률은 비교적 일정하게 유지되는 편이나, 정확률은 클러스터의 수가 증가할수록 민조 감소한다. 따라서, F-measure도 단조 감소한다. 따라서, 그림 3을 바탕으로 해서 추론해 볼 때, SOM의 셀 크기는 100 개 정도가 적당한 것으로 보인다.

5. 결 론

본 논문에서는 SOM으로 한국어 동사들을 클러스터링하는 방법을 제시하였다. 본 논문에서 제시한 방법으로 생성된 동사 클러스터들이 순수하게 정보 이론에 기반해서 만들어진 동사 클러스터들을 잘 반영하므로, 그 효용성이 입증되었다. 본 논문에서 제시한 방법은 한국어 형태소 분석기와 품사 태거 이외의 다른 정보를 사용하지 않았으므로 충분히 큰 말뭉치만 있으면 모델을 더 정확하게 확장시킬 수 있는 장점이 있다. 따라서, 한국어를 통계적 언어 모델로 분석할 때 나타나는 데이터 부족 문제는 동사에 관한한 어느 정도 줄어들 수 있으리라 생각한다.

하지만, 본 논문에서 제시한 방법은 목적어-동사 관계만을 사용하였으므로 클러스터링할 수 있는 동사가 타동사로 한정되는 문제점이 있다 이러한 문제를 해결하기 위해서 주어-동사 관계 등으로 정보를 확장하는 방법에 대한 연구를 진행할 예정이다.

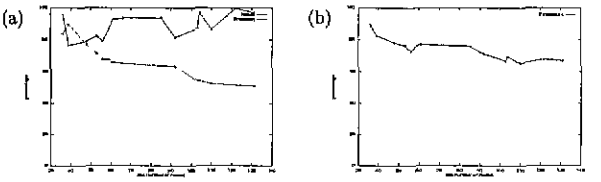

그림 3. 실험 결과. SOM으로 만든 클러스터를 상대 엔트로피 모델을 이용해 만든 클러스터와 SOM의 셀 크기를 변경해 가면서 비교하였다. (a)에서는 재현률과 정확률을 보이고 있고, (b)에서는 F-measure를 보이고 있다.

참고문헌

[1] Charniak E., *Statistical Language Learning*, MIT Press, 1993.

[2] Cover T. and Thomas J., Elements of Information Theory, Wiley-Interscience Publication, 1991.

[3] Hatzivassiloglou V and McKeown, Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning, In *Proc of the 31st Annual Meeting of the ACL*, pp 172-182, 1993.

[4] Hindle D., Noun Classification from Predicate-Argument Structures, In *Proc. of the 28th Annual Meeting of the ACL*, pp. 268-275, 1990

[5] Honkela T., Comparisons of Self-Organized Word Category Maps, In *Proc of Workshop on Self-Organizing Maps 97*, pp.298-303, 1997.

[6] Hughes J. and Atwell E., The Automated Evaluation of Inferred Word Classifications, In *Proc of the 11th European Conference on Artificial Intelligence*, pp.535-539, 1994.

[7] Jun Gao and XiXian Chen, Probabilistic Word Classification Based on a Context-Sensitive Binary Tree Method, *Computer Speech and Language*, Vol. 11, No. 4, pp 307-320, October. 1997.

[8] Pereira F., Tishby N and Lee L., Distributional Clustering of English Words, In *Proc. of the 31st Annual Meeting of ACL*, pp 183-190, 1993.

[9] Shin J. H., Han Y. S., and Choi K. S, A HMM Part-of-speech Tagger for Korean with Wordphrasal Relations, *Current Issues in Linguistic Theory. Recent Advances in NLP*, John Benjamis Publishing Company, pp.439-449, 1997.

[10] Jaehyung Yang, Conjunct Identification in Korean Noun Phrase Coordination Using Cooccurence Similarity, *Computer Processing of Oriental Language*, Vol. 10, No. 4, pp.391-408, April, 1997