단어 공기 확률 추정을 위한 차원 축소 모델

김길연, 최기선
한국과학기술원 전산학과

Dimension-Reduced Model for Word Co-occurrence Probability Estimation

Kilyoun Kim, Key-Sun Choi
Computer Science Division, KAIST
{gykim, kschoi}@world.kaist.ac.kr

요 약

본 논문에서는 확률적 자연언어 처리에서 중요한 문제인 자료 희귀(data sparseness)의 어려움을 해결하는 새로운 방법으로 차원 축소 모델을 제시한다. 세 가지의 세부 방법이 제안되었으며 Katz의 back-off 방법의 성능을 최저로 했을 때에 비해 약 60%정도의 성능이 향상되었다. 현재까지 최고의 성능을 보이고 있는 유사도 기반의 방법에 비해서도 약 5~20%의 성능이 향상되었다. 따라서 차원 축소 모델은 확률 추정의 새로운 방법으로 쓰일 수 있다.

1. 서론

확률적 자연언어 처리에서 기본 가정은 대량의 학습데이터를 기반으로 실세계의 확률분포를 추정(estimation)할 수 있다는 것이다. 이때 우리를 가장 괴롭히는 것은 자료 희귀(data sparseness)의 문제이다. 학습에 사용하는 말뭉치를 아무리 늘리더라도 실제 상황에서 미등록어는 발생하기 마련이다. 약 1,175,589개의 어절로 구성되어 있는 학습 말뭉치를 분석한 결과, 총 단어 쌍의 50%이상에 대해 bigram 확률 값을 결정할 수 없었다는 것이 보고되어 있다[1].

본 논문에서는 차원 축소를 통해 자료 희귀문제를 해결하는 새로운 방법을 제안한다. 선형 대수학의 기법인 LSA/SVD[1]를 사용하여 차원을 축소하면 자료 희귀문제를 해결할 수 있을 뿐만 아니라 단어들간에 내재된 관계까지도 파악할 수 있다.

본 논문에서 제안하는 모델은 크게 세 부분으로 나뉜다. 첫째, 추정하고자 하는 조건부 확률 분포를 행렬로 표시한다. 둘째, 생성된 행렬을 SVD 기법을 써서 낮은 차원으로 투사한다. 셋째, 축소된 차원에서 조건부 확률분포를 추정한다.

마지막 단계에서 세 가지의 추정 방법론을 제시하였으며, 이들을 기존의 추정 기법인 Katz's의 back-off 방법[2]과 유사도기반의 추정방법[3]과 비교했다. 본 모델은 back-off 모델의 성능을 최악으로 했을 경우에 비해 약 60%정도의 성능향상을 보였다.

본 논문은 다음과 같이 구성된다. 먼저 2장에서 기존의 방법들의 문제점을 살펴본다. 3장에서는 본 모델을 정의하고 4장에서 실제 실험을 통해 효용성을 검증한다. 마지막으로 5장에서 결론을 내리고 향후 연구 과제를 제시한다.

2. 관련연구

본 논문의 목적은 학습 말뭉치에 나타나지 않은 단어 쌍에 대해 조건부 확률 $P(y/x)$를 추정하는 것이다. $P(y/x)$는 단어 x가 주어졌을 때 특정 문맥 내에 단어 y가 나올 확률을 의미한다. 예를 들어, 기계 번역에서 목적어가 <맥주>로 주어졌을 때 <마시다> 혹은 <먹다>라는 동사 중에 하나를 대역어로 선택하는 경우, 확률 $P(y=마시다/x=맥주)$와 $P(y=먹다/x=맥주)$를 비교하여 확률이 큰 쪽을 선택하게 된다.

$P(y/x)$를 추정하는 가장 간단한 방법은 최

[1] LSA – Latent Semantic Analysis
SVD – Singular Value Decomposition

우추정법(Maximum Likelihood Estimation)으로 다음과 같이 주어진 단어 x의 빈도수에서 x, y가 함께 나타나는 빈도수의 비율로 정의된다.

$$P(y \mid x) = \frac{c(x,y)}{c(x)} \quad (1)$$

최우추정법의 문제점은 단어 쌍 *(x, y)*가 학습 말뭉치에 나타나지 않는 경우 확률 값이 0이 된다는 데 있다.

표1은 가상의 학습 말뭉치를 바탕으로 명사(*x*)와 동사(*y*)의 공기 확률을 최우추정법에 의해 나타낸 것으로, '음료'와 관련된 단어들과 '음식'에 관련된 단어들이 서로 구분되어 있다. 표에 의하면 *(커피, 홀짝이다)*의 단어 쌍은 학습 말뭉치에 나타나서 확률 값이 있지만, *(커피, 들이키다)*의 단어 쌍은 말뭉치에 나타나지 않아서 확률 값이 0이다. 따라서, 최우추정법에 의하면 $p(\text{들이키다}/\text{커피}) = p(\text{씹다}/\text{커피}) = 0$ 이 되어 커피에 대한 동사를 결정할 때 <들이키다>와 <씹다>를 구분할 수 없게 되는 문제가 발생한다.

이처럼 학습 데이터에 나타나지 않는 단어 쌍에 대해 확률 값이 0이 되는 문제를 해결하기 위해서 최우추정법의 수식을 조작하는 스무딩(smoothing) 방법이나 확률 값이 0이 아닌 다른 확률 정보를 혼합하여 공기 확률을 추정하는 방법들이 많이 사용되고 있다. 그 중 가장 대표적인 방법이 Katz의 back-off 방법으로 다음과 같이 정의된다[2].

$$P(y \mid x) = \begin{cases} P(y \mid x) & c(x,y) > 0 \\ \alpha P(y) & c(x,y) = 0 \end{cases} \quad (2)$$

이 방법에 의하면 빈도수가 0인 단어 쌍에 대하여 bigram 확률 $P(y/x)$를 추정함에 있어 unigram 확률인 $P(y)$만을 사용한다. 결국, 이 방법은 *x, y* 두 단어 사이의 관계를 무시하는 것으로 *y*가 같으면 *x*가 다르더라도 동일한 확률 값이 나온다는 문제가 있다. 즉, $P(\text{들이키다}/\text{커피}) = P(\text{들이키다}/\text{사탕})$ 이 되어 커피와 사탕을 구별할 수 없다.

<표1>최우추정법에 의한 단어공기확률 추정

	비우다	홀짝이다	들이키다	퍼먹다	씹다	삼키다
맥주	0.33	0	0.33	0.33	0	0
양주	0	0.5	0.5	0	0	0
커피	0	1	0	0	0	0
빵	0	0	0	0.33	0.33	0.33
사탕	0	0	0	0	0.5	0.5

이러한 전통적인 방법 외에 유사한 단어를 통해 자료 희귀 문제를 해결하는 방법(Similarity-based method)이 최근 제안되었다[3]. 이 방법은 $P(y/x) = 0$이 되는 미등록 단어 쌍 *(x, y)*에서 *x*에 가장 유사한 *k*개의 단어 *x'*를 선정한 후 이를 이용해 $P(y/x)$를 추정한다.

$$P(y \mid x) = \begin{cases} P(y \mid x) & c(x,y) > 0 \\ \frac{1}{k}\sum_{x'} P(y \mid x') & c(x,y) = 0 \end{cases} \quad (3)$$

이 방법의 문제점은 희박하게 나타나는 단어에 대해서는 유사도를 구하기가 힘들다는 점이다. 대표적인 유사도 측정 방법인 KL-Divergence[2]를 이용하면 유사도를 구하기 위해 $p(y/x)$와 $p(y/x')$의 거리를 측정한다. 그런데, 표1에서 <커피>와 <맥주>의 분포를 보면 서로 공유하는 동사가 하나도 없어서 둘 사이의 유사도는 <커피>와 <사탕> 사이의 유사도와 같게 된다. 결국 마시는 커피의 확률 $P(y/\text{커피})$을 추정하는데 있어 $P(y/\text{맥주})$ 확률 뿐만 아니라 먹는 사탕의 확률 $P(y/\text{사탕})$도 함께 이용하므로 정확도가 떨어지는 문제가 발생한다. 희박한 단어일수록 다른 단어와의 유사도를 발견하기가 힘들어지므로 자료 희귀 문제의 근본적인 해결책이 되기는 힘들다.

그러나 이러한 제반의 문제들은 본 모델에서 제안하는 차원 축소의 방법을 통해 해결이 가능하다.

3. 차원 축소 모델

이번 장에서는 본 논문에서 제안하는 차원 축소 모델에 대해서 설명한다. 차원을 축소

[2] $D(x \| x') = \sum_{y} p(y \mid x) \log \frac{p(y \mid x)}{p(y \mid x')}$

하기 위해서는 일단 조건부 확률을 행렬의 형태로 나타낼 필요가 있다. 그 후 행렬 분해 기법의 하나인 SVD를 써서 원래의 분포를 낮은 차원으로 투사한다. 이렇게 차원이 축소되면 단어들 사이에 내재된 의미적 관계가 드러나게 되고 이를 이용해 희귀한 단어에 대한 공기 확률을 추정하게 된다.

3.1 조건부 확률 행렬

모든 이산 조건부 확률 분포는 행렬로 표현할 수 있다. 확률 $p(y/x)$에 대해 주어진 단어 $x_i \in \mathbf{X}$는 열을 구성하고 예측 단어 $y_i \in \mathbf{Y}$는 행을 구성한다. 조건부 확률 행렬과 각 행, 열 벡터는 다음과 같이 정의할 수 있다. X, Y는 단어의 전체 집합이고, $m = |\mathbf{X}|$, $n = |\mathbf{Y}|$이다.

$$A_{m\times n} = [a_{ij}] = [P(y_i \mid x_i)] \quad (4)$$

$$\vec{x}_i = [P(y_1 \mid x_i), \Lambda, P(y_n \mid x_i)] \quad (5)$$

$$\vec{y}_i = [P(y_i \mid x_1), \Lambda, P(y_i \mid x_m)]$$

앞 장의 표1은 최우추정법의 수식, $A_{m\times n} = [a_{ij}] = [c(x_i, y_i)/c(x_i)]$을 적용시켜 조건부 확률 $P_{MLE}(y/x)$를 행렬로 생성한 예이다. 표1에서 각 열벡터 $\vec{x}_i = [p(y \mid x_i)]$는 6차원 즉 n=|Y|의 공간에서 표현되며, 자료 희귀의 문제가 발생하는 단어 쌍에 대해서는 각 축에서 값이 0이 된다. 이러한 0의 값들을 없애고 단어간의 내재하는 유사성을 찾기 위해 행렬 분해를 통해 n=|Y|의 차원을 보다 낮은 차원으로 줄인다.

3.2 투사 – 잠재의미분석

3.1절에서 생성한 행렬의 차원을 낮추기 위해 선형대수학의 SVD(Singular Value Decomposition)기법을 적용한다. 이 방법은 잠재의미분석(Latent Semantic Analysis)으로도 불리며, 이를 이용해 차원을 낮추게 되면 원래 행렬에서는 드러나지 않던 주어진 단어 x와 예측 단어 y간의 문맥적 연관성을 파악할 수 있다.

임의의 행렬 A에 대하여 SVD와 k-기저 근사 행렬은 다음과 같이 정의된다.

$$A = U\Sigma V^T = \sum_{i=1}^{n} u_i \cdot \sigma_i \cdot v_i^t \quad (6)$$

$$A_k = U_k \Sigma_k V_k^T = \sum_{i=1}^{n} u_i \cdot \sigma_i \cdot v_i^t \quad (7)$$

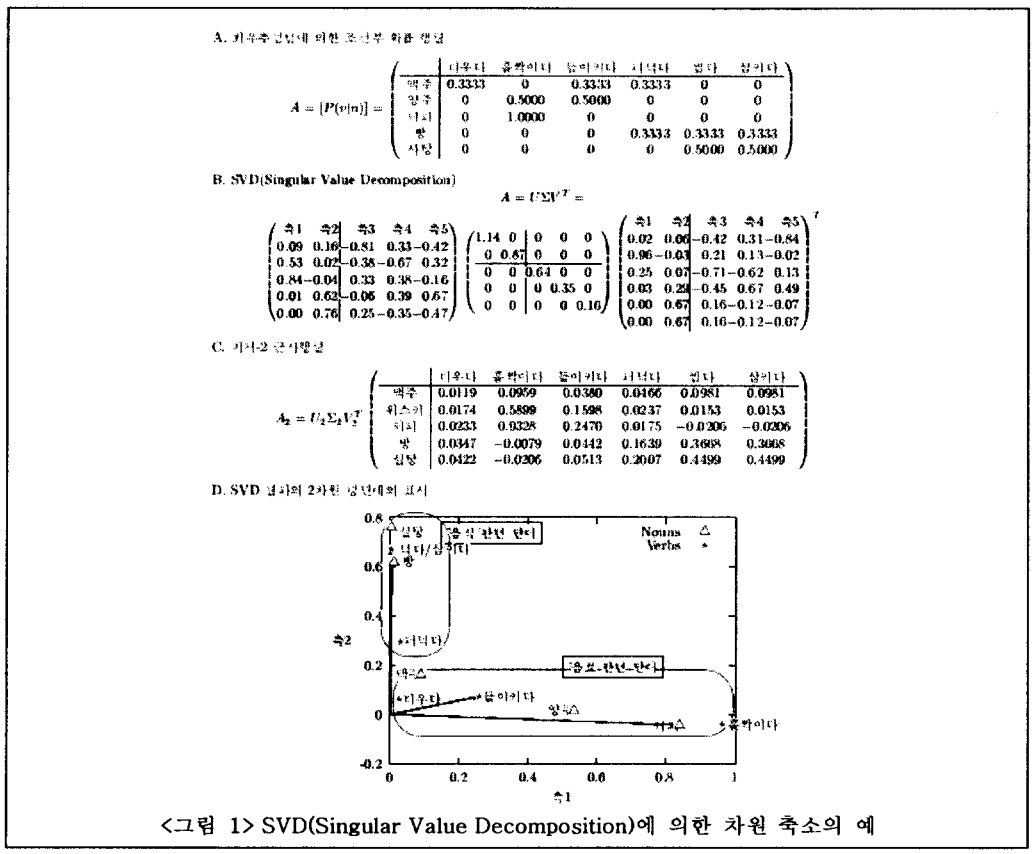
위의 식에서 Σ는 행렬 A^TA의 고유값(eigenvalue)들을 큰 순서대로 정렬한 대각 행렬이고, 좌측의 행렬 U는 AA^T에 대한 고유벡터(eigenvector)이고 V는 이와 대칭으로 A^TA에 대한 고유벡터이다. U, V는 각각 열 벡터 x와 행 벡터 y를 분해한 것이다 (그림 1-B).

한편, 각각의 행렬에서 k개의 벡터만을 취하면 원래의 행렬 A를 k차원에서 표시하면서 A에 가장 유사한 근사 행렬 기저가 k인 A_k를 구할 수 있다 (그림 1-C). 또한, U, V에서 k개의 행을 취하면 x, y를 k차원의 공간에 함께 표현할 수 있다 (그림 1-D).

그림 1은 표 1의 명사-동사의 행렬에 대해 SVD를 적용시킨 예이다. 그 중에서 그림 1-D는 원행렬의 차원을 2차원으로 줄여서 각 명사와 동사를 평면상에 나타낸 것이다. 그림에서 보듯이 차원을 줄인 결과 서로 연관성이 있는 단어끼리 군집하여 있음을 알 수 있다 (축1: '음료', 축2: '음식') 또한, <커피> 와 <맥주>는 원래의 행렬에서는 함께 공기하는 동사가 없어 유사성을 발견할 수 없었으나, 축소된 차원에서는 두 단어간의 유사도가 높다. 즉, 벡터간의 각도가 타 단어에 비해 작아서, 코사인(Cosine) 유사도가 높다. 차원을 축소한 결과 단어와 단어간의 내재된 연관성이 파악되며 이를 통해 원래 행렬에서 0이었던 $P(y/x)$ 값을 추정할 수 있는 것이다.

3.3 축소된 차원에서의 확률 추정

3.2절에서 조건부 확률의 차원을 줄였으므로 이제는 축소된 차원에서 $P(y/x)$값을 어떻게 추정할 수 있는지



<그림 1> SVD(Singular Value Decomposition)에 의한 차원 축소의 예

살펴본다. 총 3가지의 방법을 생각해 볼 수 있다.

3.3.1 제 1 방법: 거리 기반의 추정

식 (6, 7)에서 SVD를 통해 원행렬 A는 U, Σ, V 의 세 행렬로 분해되고 U, V행렬에서 k 개의 행만을 취하면 단어 x와 y를 k차원에 표시할 수 있다 (그림 1-D). 거리 기반의 추정에서는 축소된 차원에서 단어 x, y간의 거리를 구하는 것으로, 행, 열 벡터 $\mathbf{x}$, $\mathbf{y}$를 k차원으로 투사한 결과인 $\mathbf{u}$, $\mathbf{v}$ 벡터 간의 거리를 구한다.

$$p(y_i \mid x_i) = \frac{1}{Z_k} \frac{D_k(\vec{u}_i, \vec{v}_i) + 1}{2},$$

$$D_k(\vec{u}_i, \vec{v}_j) = \frac{\sum_{t=1}^{k} u_i(t) v_j(t)}{\sqrt{\sum_{t=1}^{k} u_i(t)^2} \sqrt{\sum_{t=1}^{k} v_i(t)^2}} \quad (8)$$

3.3.2 제 2 방법: k-기저 근사행렬 기반의 추정

식 (6)에서 SVD 결과인 U, E, V벡터에서 k개의 행만을 취해 곱하면 식 (7)과 같이 k 차원에서 A에 가장 근사한 행렬인 A_k 행렬을 구할 수 있다 (그림 1-C). 이 방법에서는 A_k 행렬의 각 원소를 $P(y/x)$로 간주한다. 단, A_k 의 각 원소에서 음수값을 제거하고 확률로 만들기 위해 다음과 같이 정의한다. Z_k는 정규화를 위한 변수이고 δ는 스무딩 (smoothing) 상수이다.

$$p(y_i \mid x_i) = \frac{1}{Z_k}\left[A_k(i,j) - \arg\min_y A_k(i,j) + \delta \right],$$

$$Z_k = \sum_{j=1}^{n}\left(A_k(i,j) - \arg\min_y A_k(i,j) + \delta \right) \quad (9)$$

3.3.3 제 3 방법론: 축소된 차원에서의 유사도 기반의 추정

자료 희귀의 문제를 해결하는데 현재까지 가장 좋은 결과를 보이고 있는 유사도 기반의 방법[3]과 본 논문에서 제안하는 차원 축소 모델은 하나로 합쳐질 수 있다. 즉, 축소된 차원에서 단어간의 유사도를 계산하여 k 개의 가장 유사한 단어들을 추정에 사용하는 것이다. 다음과 같이 정의된다.

$$P(y_j|x_i)=\begin{cases} P_{MLE}(y_j|x_i) & c(x,y)>0 \\ \frac{1}{|S(x_i,k,\vartheta)|}\sum_{x_i'\in S(x_i,k,\vartheta)} P_{MLE}(y_j|x_i') & c(x,y)=0 \end{cases}$$

원래 유사도 기반의 모델에서는 JS-divergence를 유사도 측정치로 사용했으나 [4] 여기에서는 코사인(cosine)을 사용한다. 또한, 가장 유사한 k개의 단어는 임계값 θ로 결정된다.

4. 실험

본 모델의 효용성을 검증하기 위해 유사도 기반의 모델에서 사용된 방법과 동일하게 실험을 수행했다[5]. 명사가 주어지고 그 명사를 목적어로 취하는 동사의 확률을 추정하는 실험이다. 전체 말뭉치에서 명사 n에 대해 두 개의 동사 v_1, v_2를 제시하되 $c(n,v1) > 2 * c(n,v2)$ 가 되도록 $v1$, $v2$를 선택했다. 이후 학습집합과 시험집합으로 말뭉치를 나눌 때에 학습집합에서 $(n,v1)$, $(n,v2)$의 단어쌍을 제외하고 이를 실험에 이용했다. 따라서, 만약 $P(v1|n) > P(v2|n)$의 결과가 나오면 정답이고 반대의 결과가 나오면 오류이다. 성능은 다음과 같은 오류율로 평가한다.

$$오류율 = \frac{실제오류의수}{총시험집합의크기}$$

한편, 전통적인 방법이면서 가장 널리 쓰이는 Katz의 back-off 방법을 실험의 기준으로 삼기 위해 $c(v1) < c(v2)$가 되도록 $v1$, $v2$를 선택하여 back-off의 경우 오류율이 100%가 되도록 하였다 (식 (2) 참조). 실험은 3-fold 교차검증(Cross-validation)으로 수행했다. 아래 표 2는 데이터의 준비 科程을 요약한 것이다.

<표 2> 학습집합과 시험집합

0. 말뭉치	Penn Treebank II
1. 동사-목적어 단어 쌍 추출	18843 쌍
2. 명사×동사 (행렬 크기)	1000 × 2008
3. 학습 집합 평균 크기	13040 쌍
4. 실험 집합 평균 크기	713 쌍

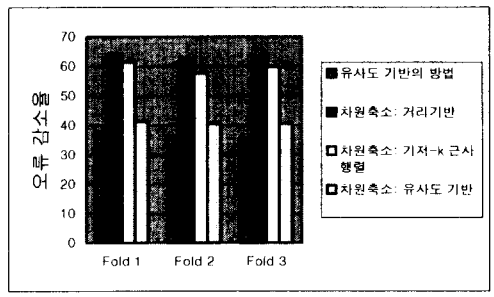
4.1 실험결과

표 3과 그림 2는 3-fold의 시험집합에 대한 실험 결과를 평균한 것이다. 표 3은 오류율을, 그림 2는 back-off에 비교한 오류 감소율을 나타낸다. 방법론 1, 2가 가장 좋은 성능을 보이고 있으며, Katz's back-off방법의 오류율을 100%로 했을 때 평균적인 성능 향상은 60%정도이다. 방법론 3은 방법론 1, 2에 비해서는 성능이 떨어지지만 기존의 유사도 모델에 비해서는 더 나은 결과를 보인다.

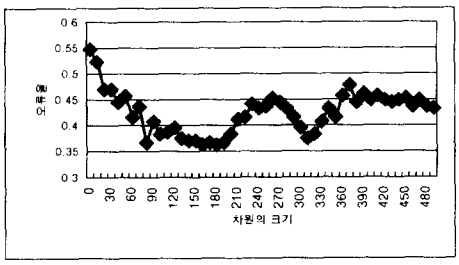
이처럼 차원 축소 모델이 좋은 성능을 보이는 이유는 SVD에 의해 각 열간의 단어 유사도 뿐만 아니라 각 행간의 단어 유사도가 함께 확률 추정에 반영되기 때문이다.

<표 3> 실험 결과

	Back-off	유사도 기반	차원 감소 모델		
			거리기반	기저k 행렬	유사도 기반
Fold1	1.0	0.623	0.362	0.386	0.586
Fold2	1.0	0.636	0.374	0.423	0.594
Fold3	1.0	0.645	0.366	0.402	0.593



<그림 2> 에러 감소율



<그림 3> 최적 차원의 수

한편, 그림 3은 차원의 크기를 줄여 나감에 따라 오류율이 어떻게 변하는 지를 나타내는 그래프이다. 원행렬의 크기는 1000이었으며 축소한 차원의 크기가 90~200사이에서 최고의 성능을 보였다. 이는 곧 1000차원에서 나타나지 않았던 단어간의 내재된 연관성이 아주 낮은 차원에서 잘 드러난다는 사실을 말해준다.

5. 결론

본 논문에서는 확률적 자연언어 처리에서 중요한 문제인 자료 희귀(data sparseness)의 어려움을 해결하는 새로운 방법으로 차원 축소 모델을 제시했다. 세 가지의 세부 방법이 제안되었으며 Katz의 back-off 방법의 성능을 최저로 했을 때에 비해 약 60%정도의 성능이 향상되었다. 현재까지 최고의 성능을 보이고 있는 유사도 기반의 방법에 비해서도 약 5~20%의 성능이 향상되었다. 따라서 차원 축소 모델은 확률 추정의 새로운 방법으로 쓰일 수 있다.

차원 축소에 쓰인 SVD기법을 이용하면 단어들간의 내재된 관계를 효과적으로 뽑아낼 수 있다. 이 방법은 수학적으로 완전히 자동화된 과정이므로 일단 차원을 축소해 놓으면 별도의 추가비용 없이 확률 추정이 가능하다. 또한, Folding-in, updating의 방법으로 이미 축소된 차원에 단어를 동적으로 투사하는 것도 가능하다[5].

한편, 가장 최적의 성능을 보이는 차원의 크기를 이론적으로 결정하는 문제를 차후에 해결해야 한다[6]. 본 실험에서는 원행렬의 차원이 1000일때 90~200 차원 정도에서 최상의 성능을 보였는데, 실험적으로 이를 찾는 비용이 상당하다. 덧붙여, 빈도 정보가 아닌 다른 정보를 함께 포함시킬 수 있는 보다 일반적인 틀에 대한 연구도 필요하다.

참고 문헌

[1] 강인호. 1999. 최대 엔트로피 모델을 이용한 한국어 품사 태깅. 한국과학기술원 전산학과, 석사학위논문

[2] Slava M. Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-35(3):400-401.

[3] Ido Dagan, Lillian Lee, and Fernando C. N. Pereira. 1999. Similarity-based models of word co-occurrence probabilities. *Machine Learning*, 34:43-69

[4] Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 25-32

[5] Michael W. B., Zlatko D., and Elizabeth R. J. (1999). Matrices, Vector Spaces, and Information Retrieval. *Society for Industrial and Applied Mathematics, SIAM Review Vol 41, No. 2*, pp. 335-362

[6] Chris H.Q Ding. (1999). A Similarity-based Probability Model for Latent Semantic Indexing. In *22nd Annual International ACM SIGIR Conference (SIGIR '99)*, pp. 58-65