언어모델의 데이터 희소성 문제 개선을 위한 앙상블 기법

장하영 O 장병탁

서울대학교 공과대학 컴퓨터공학부

{hyjang, kmkim}@bi.snu.ac.kr {timothy.park, shwang}@samsung.com btzhang@bi.snu.ac.kr

Ensemble Methods for Data Sparseness Problem in Language Model

Ha-Young Jang O Byoung-Tak Zhang

Department of Computer Science and Engineering, Seoul National University

요 약

언어 데이터의 특성으로 인해서 발생하는 데이터의 희소성 문제는 일반적으로 사용되는 n-gram 기반의 언어모델에서 말뭉치에 존재하는 단어의 종류가 증가함에 따라서 지수적으로 커지게 된다. 이의 해결을 위하여 다양한 종류의 평활법(smoothing)을 사용함으로써 언어모델의 성능을 향상시키는 것이 가능하다. 본 논문에서는 이러한 데이터의 희소성 문제를 개선하기 위한 하이퍼그래프 기반의 앙상블 언어모델을 제안하였다. 제안한 모델은 앙상블 기법의 적용으로 인해서 희소성 문제를 감소시킬 수 있을 뿐만 아니라 적절한 평활법을 함께 이용함으로써 일반적인 n-gram 언어모델보다 성능을 크게 향상시킬 수 있다. 본 논문에서는 이러한 결과를 학습된 언어모델의 혼잡도(perplexity)를 이용하여 평가하였다.

1. 서 론

문자열의 확률을 추정하기 위해서 사용되는 n-gram 언어모델은 최대 우도 추정(maximum likelihood estimation)을 이용하여 확률을 계산한다. 언어모델의 학습을 위해서 사용한 말뭉치에 존재하는 단어들의 집합을 V라 할 때, n=3이고 $|V|$=1,000이라고 하면 총 계산해야 하는 확률의 개수는 $|V|^3=10^{12}$개가 된다. 이러한 경우에 일반적인 크기의 말뭉치에서는 최대 우도 추정을 통해 확률값을 계산할 경우에 대부분의 확률이 0의 값을 가지게 되는데, 이러한 문제를 해결하기 위해서 다양한 평활법들이 연구되었다[1]. 이러한 평활법의 적용은 언어 데이터에 존재하는 희소성 문제를 개선시켜서 n-gram 언어모델의 성능 향상에 많은 도움이 되고 있다.

희소성 문제의 해결을 위한 앙상블 기법의 대표적인 예로 랜덤 포레스트(random forest)기반의 언어모델이 있다[2]. 랜덤 포레스트 언어모델에서는 일반적인 앙상블 기법에서와 마찬가지로 표본화(sampling)을 통해서 만들어진 의사결정나무(decision tree) 언어모델의 앙상블로 이루어진 랜덤 포레스트를 이용하여 문자열의 확률을 추정하게 된다. 랜덤 포레스트 언어모델의 경우에 있어서도 n-gram 언어모델의 경우와 마찬가지로 앙상블 모델의 구축과정에서 평활법의 사용이 필요하지만, 데이터 희소성 문제를 관측되지 않은 사례에 대해서도 일반화가 가능한 모델을 찾는 문제라고 생각할 때 구축된 모델의 일반화에 유리한 앙상블 기법의 본질적인 특성이 문제 해결에 많은 도움이 될 수 있을 것으로 판단된다. 본 논문에서 제안한 방법은 이러한 앙상블 모델의 특성을 이용하고 있지만, 랜덤 포레스트 언어모델과는 다르게 일반적인 n-gram 언어모델에서 사용되는 n-gram과 유사한 역할을 하는 하이퍼에지를 이용하여 앙상블 모델을 구축한다는 차이점이 있다.

본 논문의 구성은 다음과 같다. 2장에서는 하이퍼그래프 언어모델에 대해서 논의하고, 3장에서는 이를 이용한 앙상블 언어모델에 대해서 설명하겠다. 이 후 4장에서 희소성이 큰 말뭉치를 이용하여 제안한 모델의 성능을 확인한 결과를 제시한다. 이 후 5장에서 결론 및 향후 연구 방향을 모색한다.

2. 하이퍼그래프 언어모델

n-gram 언어모델에서 n-gram을 이용하여 최대 우도

추정을 통하여 확률값을 구하는것과 유사하게 하이퍼그래프 모델에서는 말뭉치에 존재하는 단어들의 조합으로 구성된 하이퍼에지와 하이퍼에지의 출현빈도를 표현하는 가중치로 확률분포를 표현하게 되는데 이때 하이퍼그래프 모델의 에너지는 다음과 같이 정의된다[3].

$$\varepsilon\left(s^{(n)};WGT\right) = -\sum_{i=1}^{|E|} wgt_{i_1 i_2 \dots i_{|E_i|}} x_{i_1}^{(n)} x_{i_2}^{(n)} \dots x_{i_{|E_i|}}^{(n)} \quad (1)$$

$x_{i_1}^{(n)} x_{i_2}^{(n)} \dots x_{i_{|E_i|}}^{(n)}$ 는 말뭉치 내에 존재하는 단어들의 조합으로 구성된 하이퍼에지를 의미하고 *WGT*는 하이퍼에지의 가중치를 의미한다. 즉, n-gram 방식의 언어모델에서 사용하는 단어들의 출현빈도를 가중치의 형태로 표현하여 이를 이용하여 하이퍼그래프 모델을 표현하는 것이다.

이렇게 만들어진 하이퍼그래프 모델에서 문장 $s^{(n)}$이 나타날 확률은 다음과 같이 깁스 분포에 의해서 주어지게 되고,

$$P\left(s^{(n)}\middle|WGT\right) = \frac{1}{Z(WGT)} exp\left\{-\varepsilon(s^{(n)};WGT\right\} \quad (2)$$

분할함수(partition function) *Z(WGT)*는 다음과 같이 정의된다.

$$Z(WGT) = \sum_{x^{(m)}} exp\left\{-\varepsilon(s^{(m)};WGT\right\} \quad (3)$$

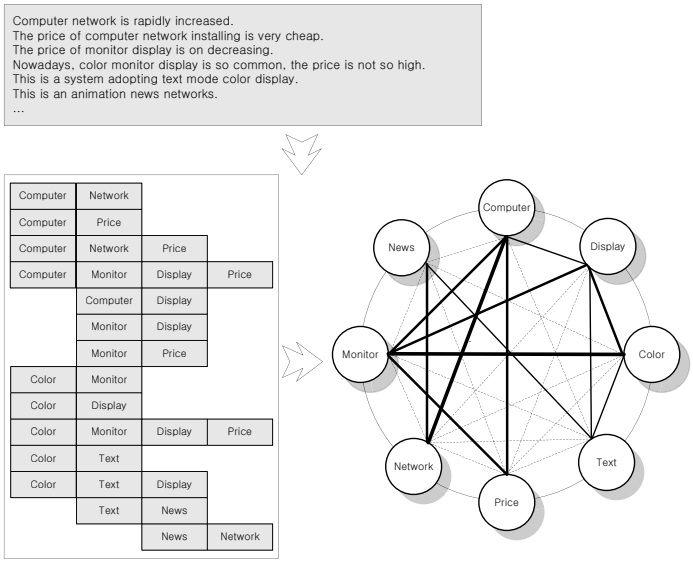

그림 1. 하이퍼그래프 모델을 이용한 언어모델 구축 예.

3. 하이퍼그래프 기반 앙상블 언어모델

하이퍼그래프 기반 앙상블 언어모델에서 문자열의 확률은 아래와 같이 하이퍼그래프 언어모델을 이용하여 정의할 수 있다.

$$P\left(w_i\middle|w_{i-n+1}^{i-1};H_k\right),\ k=1,\dots \quad (4)$$

이 때, H_k 는 각각의 하이퍼그래프 언어모델을 나타내는 확률변수이고 각각의 하이퍼그래프 언어모델의 확률을 앙상블한 결과는 아래와 같다.

$$P\left(w_i\middle|w_{i-n+1}^{i-1}\right) = \frac{1}{M}\sum_{j=1}^{M} P_{H_j}\left(w_i\middle|\Phi_{H_j}\left(w_{i-n+1}^{i-1}\right)\right) \quad (5)$$

이 때, $\Phi_{H_j}\left(w_{i-n+1}^{i-1}\right)$는 문자열 w_{i-n+1}^{i-1}를 하이퍼그래프 언어모델에 사상시키기 위한 사상함수(mapping function)이다. 그러나 미관측 데이터에 대해서는 사상함수가 정의되지 않기 때문에, 이를 해결하기 위해서 랜덤 포레스트 언어모델에서 사용한 interpolated Kneser-Ney 평활법을 이용하였다.

Interpolated Kneser-Ney 평활법은 아래와 같이 정의된다.

$$P_{KN}\left(w_i\middle|w_{i-n+1}^{i-1}\right) = \frac{\max\left(C\left(w_{i-n+1}^{i}\right)-D,0\right)}{C\left(w_{i-n+1}^{i-1}\right)} + \lambda\left(w_{i-n+1}^{i-1}\right)P_{KN}\left(w_i\middle|w_{i-n+2}^{i-1}\right) \quad (6)$$

이 때, D는 절단 상수(discounting constant)이고 $\lambda\left(w_{i-n+1}^{i-1}\right)$ 는 보간 가중치(interpolation weight)이다[4]. 하이퍼그래프 언어모델에서 문자열의 확률을 구하기 위하여 이를 적용하면 아래와 같다[5].

$$P_H\left(w_i\middle|\Phi\left(w_{i-n+1}^{i-1}\right)\right) = \frac{\max\left(C\left(w_i,\Phi\left(w_{i-n+1}^{i-1}\right)\right)-D,0\right)}{C\left(\Phi\left(w_{i-n+1}^{i-1}\right)\right)} + \lambda\left(\Phi\left(w_{i-n+1}^{i-1}\right)\right)P_{KN}\left(w_i\middle|w_{i-n+2}^{i-1}\right) \quad (7)$$

따라서 식 (5)와 식 (7)을 이용하여 앙상블 언어모델에서 문자열의 확률을 계산할 수 있다.

4. 실험 및 결과

본 논문에서 제안한 방법이 희소성 문제의 개선에 효과적인지 확인하기 위하여 본 실험에서는 대용량의 말뭉치를 사용하지 않고, 희소성이 잘 드러날 수 있는 데이터를 사용하였다. 이를 위해서 데이터의 분량이

그다지 크지 않고 다양한 주제로 인해서 희소성이 더욱 강한 특성을 갖는 시트콤 'Friends'의 대사로 구성된 말뭉치를 선택하여 실험을 진행하였다. 말뭉치는 전체 10 개 시즌, 223 개의 에피소드에 나오는 모든 대사로 구성이 되어 있고, 교착어인 한국어의 특성으로 인해서 발생하는 어휘의 증가와 전처리 과정에서 발생할 수 있는 오류를 감소시키기 위해서 영어 대사를 이용하여 실험을 진행하였고, n-gram 에서의 n 과 같은 역할을 하는 하이퍼에지의 차수는 3 을 사용하였다.

학습된 언어모델의 평가를 위해서 아래와 같이 정의되는 혼잡도(perplexity)를 척도로 사용하였다[1].

$$PPL(M) = exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log(P(w_i|w_1^{i-1}))\right) \quad (8)$$

이 때, $w_1, \ldots, w_N$ 은 N 개의 단어로 이루어진 문자열이다.

첫번째 실험은 학습에 사용한 말뭉치를 얼마나 잘 모델링하는지 확인하기 위하여 평활법을 적용한 일반적인 n-gram 언어모델을 성능을 비교해 보았다. 표 1 에서 볼 수 있는 것처럼 제안한 방법론이 더 낮은 혼잡도를 보이고 있고, 보다 효과적으로 학습에 사용한 말뭉치의 특성을 모델링하고 있다고 판단된다.

Model	Perplexity
Trigram with Kneser-Nay	221.3
Ensemble Language Model	198.2

표 1. n-gram 모델과의 성능 비교

두번째 실험에서는 제안한 모델의 일반화 성능을 확인하기 위해서 총 10 개의 시즌 중에서 하나의 시즌을 제외한 나머지 말뭉치로 언어모델을 구축한 후에 언어모델 구축에 사용되지 않은 다른 하나의 시즌에 대한 혼잡도를 측정해 보았다. 표 2 에 총 10 회의 실험에서 측정한 평균 혼잡도가 나와 있다. 각각의 실험에서는 서로 다른 시즌을 테스트 말뭉치로 사용하였다.

Model	Perplexity
Trigram with Kneser-Nay	449.0
Ensemble Language Model	362.8

표 2. 미관측 데이터에 대한 성능 비교

표 2 의 실험 결과에서 확인할 수 있듯이 제안한 방법론이 미관측 데이터에 대한 성능에서도 일반적인 n-gram 언어모델보다 우수한 성능을 보이고 있음을 확인할 수 있다.

4. 결론

본 논문에서는 언어 데이터에서 발생하는 희소성 문제를 개선하기 위한 앙상블 기반의 언어모델을 제안하였다. 앙상블 기법은 데이터의 표본화를 통해서 만들어진 다양한 모델을 결합하여 최종 모델을 구축하는 특성으로 인해서 일반적인 경우에 모델의 일반화 성능이 매우 우수한 특성을 보인다. 이러한 특성은 희소성 문제가 관측되지 않은 데이터에 대한 일반화의 부족으로 인해서 발생한다는 점을 감안할 때 희소성 문제에 많은 도움을 줄 수 있을 것이라 쉽게 예상할 수 있고, 앙상블 기법을 적용한 제안 모델의 실험 결과에서도 이를 확인할 수 있다.

현재 제안한 방법론에서는 모델의 성능 향상을 위해서 일반적인 평활법의 개념을 적용할 수 밖에 없는 한계점이 존재하는데, 이러한 부분들을 랜덤 포레스트에서 이용하는 특성 벡터(feature vector)의 표본화 등을 적용함으로써 개선할 수 있으리라 기대된다.

감사의 글

이 논문은 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구(NRF-2010-0017734-Videome,)임.

참고문헌

[1] Simonoff, J. S., Smoothing methods in statistics, *Springer*, 1996.

[2] Xu, P. and Jelinek, F., Random forests in language modeling, Random forests and the data sparseness problem in language modeling, *Computer Speech and Language*, 21(1):105-152, 2007.

[3] Zhang, B.-T., Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory, *IEEE Computational Intelligence Magazine*, 3(3) pp. 49-63, 2008.

[4] Kneser, R., Ney, H., Improved backing-off for n-gram language modeling. *In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 181–184, 1995.

[5] Chen, S.F., Goodman, J., An empirical study of smoothing techniques for language modeling. *Tech. Rep. TR-10-98, Computer Science Group, Harvard University, Cambridge, MA*., 1998