김일성종합대학학보
(자연과학)
주체106(2017)년 제63권 제4호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 63 No. 4 JUCHE106(2017).

# 조선어련속음성인식을 위한 대규모재귀신경망언어모형 구축의 한가지 방법

**리현순, 문성일**

최근 음성인식체계들에서는 지금까지 많이 리용되여오던 통계적N－그람언어모형의 제한성을 극복하기 위하여 신경망을 언어모형학습에 리용하고있다.[2, 3]

재귀신경망언어모형(RNNLM)은 단어리력을 제한하지 않고 임의의 길이를 가지는 단어문맥을 리력으로 리용하여 련속공간에서 파라메터추정을 진행함으로써 신경망언어모형의 성능을 높이고있다.[2]

현재 어휘규모가 크지 않은 조선어련속음성인식체계에서 RNNLM을 리용하여 인식률을 개선[1]하고있으나 대규모학습자료를 리용하는 경우 숨은층의 크기를 증가시키는데 따라 계산량이 늘어나고 학습속도가 떨어져 대어휘련속음성인식체계들에는 아직까지 도입되지 못하고있다.

론문에서는 음성인식체계의 대규모언어모형학습에 재귀신경망을 리용하는데서 나서는 학습속도개선방법을 제안하였다.

## 1. 출력층분해를 리용한 재귀신경망언어모형구축방법

### 1) 재귀신경망의 출력층분해

재귀신경망 출력층의 어휘들을 무리짓기하여 클라스나무를 생성하는 방법으로 출력층을 분해한다.

클라스나무에서 매 단어는 하나의 클라스와 그것과 련관된 부분클라스들에 속한다.

$w_i$가 문장에서 $i$번째 단어라고 하면 렬 $c_{1,D}(w_i)=c_1(w_i),\cdots,c_D(w_i)$는 클라스나무에서 단어 $w_i$에 대한 경로를 반영한다. 여기서 $D$는 나무의 깊이, $c_d(w_i)$ ( $d=\overline{1,\ D-1}$ )는 $w_i$에 대응하는 클라스 또는 부분클라스, $c_D(w_i)$는 $w_i$와 련관된 잎(단어)을 표현한다.

단어리력 $h$가 주어진 조건에서 $w_i$의 N－그람확률은 다음과 같이 추정된다.

$$P(w_i|h)=P(c_1(w_i)|h)\prod_{d=2}^{D}P(c_d(w_i)|h,\ c_{1,\ d-1}) \tag{1}$$

재귀신경망의 출력층을 3개의 부분층들로 분해한다.

첫번째 출력층은 단축목록(shortlist)단어들과 OOS단어(단축목록밖의 단어)들을 위한 상위클라스들(가장 일반적인 클라스들)의 분포를 추정한다.

단축목록에 있는 단어들은 매개가 부분클라스가 없이 자기자체의 클라스를 표현한다. 이 경우에는 $D=1$이다. 상위클라스들은 OOS단어들을 처리하기 위한 나무의 뿌리로 되며 매 층이 softmax함수를 가지는 여러개의 부분클라스층들을 포함한다. 이 부분클라스층들이 두번째 층을 형성한다.

세번째 층인 단어층들은 OOS단어들에 대한 단어확률을 추정한다.

식 (1)에서 상위클라스들에 대한 분포 $P(c_1(w_i)|h)$를 계산하기 위해 첫번째 출력층이 리용된다. 그리고 부분클라스층들은 $1<d<D$에 대하여 부분클라스확률 $P(c_d(w_i)|h, c_{1,\ d-1})$들을 계산하는데 리용된다. 마지막으로 OOS단어확률 $p(C_D(w_i)|h, c_{1,\ D-1})$들은 단어층에서 계산된다.

출력층분해를 리용한 재귀신경망의 구조는 그림과 같다.

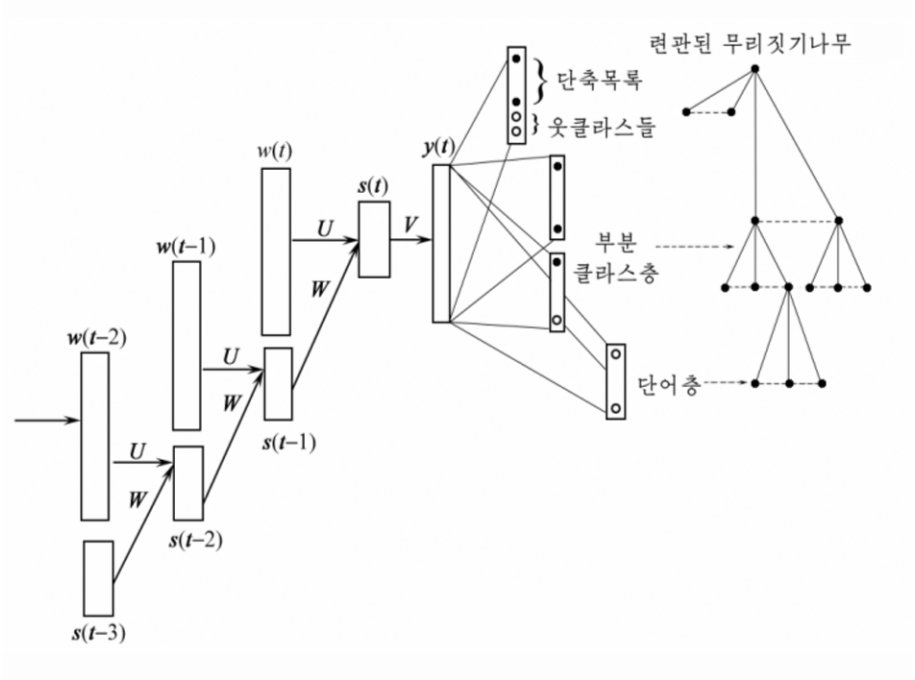


그림. 출력층분해를 리용한 재귀신경망의 구조

망은 입력층 $x$와 숨은층 $s$, 출력층 $y$를 가진다.

입력층과 숨은층사이에는 무게행렬 $\boldsymbol{U}$와 $\boldsymbol{W}$, 숨은층과 출력층사이에는 무게행렬 $\boldsymbol{V}$가 있다. 시각 $t$에서 망의 입력은 $\boldsymbol{x}(t)$이고 출력은 $\boldsymbol{y}(t)$이며 $\boldsymbol{s}(t)$는 숨은층의 상태이다. 입력벡토르 $\boldsymbol{x}(t)$는 현재단어를 표현하는 벡토르 $\boldsymbol{w}(t)$와 시각 $t-1$에서 숨은층의 출력 $\boldsymbol{s}(t-1)$로부터 얻어진다. 출력층은 클라스나무를 리용하여 3개 부분층으로 분해되였다.

숨은층, 출력층들은 각각 다음과 같이 계산된다.

$$s_j(t) = f\left(\sum_i w_i(t)u_{ji} + \sum_l s_l(t-1)w_{jl}\right) \tag{2}$$

$$y_k(t) = g\left(\sum_j s_j(t)v_{kj}\right) \tag{3}$$

여기서 $f(z)$는 시그모이드활성함수로서

$$f(z) = 1/(1+e^{-z}) \tag{4}$$

이며 $g(z)$는 softmax함수이다. 즉

$$g(z_m) = e^{Z_m}\Big/\sum_k e^{Z_k}\ . \tag{5}$$

2) 출력층분해를 리용한 재귀신경망언어모형의 학습

이 학습은 출력층을 클라스나무로 분할하는 단어무리짓기단계와 전체 어휘를 가진 신경망파라메터학습단계로 나누어 진행한다.

알고리듬은 다음과 같다.

단계 1 클라스나무생성을 위한 단어무리짓기

① 단어특징학습

신경망의 출력으로 단축목록만을 가지는 RNN을 학습하여 무게행렬 $\boldsymbol{U}$와 $\boldsymbol{W}$를 추정한다.

② 사영공간의 차원축소

행렬 $\boldsymbol{U}$에 대하여 표준주성분분석(PCA)방법을 적용한다.

③ 클라스나무생성

단축목록에 없는 단어들을 가지고 앞선단계에서 생성된 단어특징벡토르에 기초하여 K-평균알고리듬으로 하강형단어무리짓기를 진행한다. 클라스에 속하는 단어들의 개수가 실험적으로 설정된 턱값 $W$보다 큰 경우에는 부분클라스들로 가른다. $W$개이상의 단어들을 포함하는 매 클라스는 $|\sqrt{W}+1|$개의 부분클라스들로 나누어진다.

단계 2 전체 어휘를 가진 재귀신경망학습

출력으로서 클라스나무구조를 가지는 전체 어휘 RNNLM을 BPTT알고리듬을 리용하여 학습한다.

한주기 RNNLM학습은 다음과 같이 진행된다.

① 시각 $t$를 0으로 설정하고 숨은층에서 신경세포들의 상태 $\boldsymbol{s}(t)$를 초기화한다.

② $t$를 1만큼 증가시킨다.

③ 입력층에 현재단어 $w_t$를 표현하는 단어벡토르 $\boldsymbol{w}(t)$를 입력한다.

④ 숨은층의 상태 $\boldsymbol{s}(t-1)$를 입력층에로 복사한다.

⑤ 식 (2), (3)에 따라 $\boldsymbol{s}(t)$, $\boldsymbol{y}(t)$를 계산한다.

⑥ 출력층에서 오차 $\boldsymbol{e}(t)$의 그라디엔트를 교차엔트로피척도를 리용하여 계산한다.

⑦ 오차를 역전파시키고 무게들을 대응하게 변화시킨다.

$$v_{jk}(t+1)=v_{jk}(t)+s_j(t)e_{ok}(t)\alpha \tag{6}$$

여기서 $\alpha$는 학습률, $j$는 숨은층의 크기, $k$는 출력층의 크기, $s_j(t)$는 숨은층에서 $j$번째 신경세포의 출력이다. 그리고 $e_{ok}(t)$는 출력층에서 $k$번째 신경세포의 오차그라디엔트이다. 오차그라디엔트는 숨은층 $\boldsymbol{s}(t)$로부터 이전숨은층 $\boldsymbol{s}(t-1)$에로 재귀적으로 전파되는데 이것은 성분별로 적용되는 함수 $d_h(\cdot)$를 리용하여 다음과 같이 얻는다.

$$\boldsymbol{e}_h(t-\tau-1)=d_h(\boldsymbol{e}_h(t-\tau)^{\mathrm{T}}\boldsymbol{W},\ \ t-\tau-1). \tag{7}$$

여기서 $\tau$는 시간지연을 나타내며 $d_{hj}(x,\ t)=xs_j(t)(1-s_j(t))$이다.

이때 입력층 $\boldsymbol{w}(t)$와 숨은층 $\boldsymbol{s}(t)$사이의 무게행렬 $\boldsymbol{U}$는 다음과 같이 갱신된다.

$$u_{ij}(t+1)=u_{ij}(t)+\sum_{z=0}^{T}w_i(t-z)e_{hj}(t-z)\alpha-u_{ij}(t)\beta \tag{8}$$

여기서 $T$는 시각의 개수이다.

한편 재귀무게행렬 $\boldsymbol{W}$는 다음과 같이 갱신된다.

$$w_{lj}(t+1) = w_{lj}(t) + \sum s_l(t-z-1)e_{hj}(t-z)\alpha - w_{lj}(t)\beta \quad (9)$$

⑧ 모든 학습표본들을 처리하지 못했으면 단계 2로 간다.

## 2. 성 능 평 가

조선어음성인식프로그람《룡남산》을 리용하여 대규모RNNLM의 성능평가실험을 진행하였다. 실험에서는 우선 대규모재귀신경망언어모형의 출력층분해를 리용한 학습속도개선평가 및 최량인 단축목록크기를 결정하였다. 대비기준모형으로서는 출력층분해가 없는 재귀신경망언어모형을 리용하였다. 언어모형학습자료로서 45M단어들로 구성되는 본문자료, 평가자료로서 10M단어들로 구성되는 본문자료(학습에 참가하지 않은)를 리용하였으며 생성된 어휘수는 60K이다.

실험은 Core i3(2GB, 3GHz)급콤퓨터에서 숨은층의 크기를 150으로 고정하고 진행하였다. RNNLM 1은 출력층분해가 없는 재귀신경망언어모형, RNNLM 2는 출력층분해를 리용한 재귀신경망언어모형이라고 하자. RNNLM 1에서는 한주기 학습하는데 보통 38h, 10주기 학습하는데 15일이상의 시간이 걸린다. 한편 RNNLM 2에서는 단축목록의 크기에 따라 한주기 학습하는데 보통 3～13h정도 걸린다.(표 1)

표 1. 단축목록의 크기에 따르는 한주기 학습시간(*T*)과 분기수(PPL)

| 단축목록크기 | 500 | | 1 000 | | 1 500 | | 2 000 | | 2 500 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *T*/h | PPL | *T*/h | PPL | *T*/h | PPL | *T*/h | PPL | *T*/h | PPL |
| RNNLM 1 | 38 | — | 38 | — | 38 | — | 38 | — | 38 | — |
| RNNLM 2 | 3 | 136 | 4 | 128 | 5.5 | 113 | 8 | 109 | 13 | 109 |

실험을 통하여 최량인 단축목록크기는 2 000으로 설정하였으며 이때 학습속도는 4.7배 개선되였다는것을 알수 있다.

실험에서는 다음으로 제안된 방법의 성능평가를 20K규모의 어휘크기에서 진행하였다.(표 2)

실험을 통하여 RNNLM 1에 비하여 제안한 방법의 성능이 약간 떨어진다는것을 알수 있다.

표 2. 제안된 방법의 성능평가

| 모형 | 분기수(PPL) | 단어오유률(WER)/% |
|---|---|---|
| RNNLM 1 | 85 | 2.31 |
| RNNLM 2 | 87 | 2.33 |

## 맺 는 말

재귀신경망언어모형학습에 출력층분해를 적용하여 성능을 거의 유지하면서도 학습속도를 훨씬 개선함으로써 조선어련속음성인식을 위한 대규모재귀신경망언어모형학습에 리용할수 있게 하였다.

## 참 고 문 헌

[1] 리현순; 정보과학, 2, 52, 주체104(2015).

[2] T. Mikolov et al.; Proceedings of ICASSP, 11, 5528, 2011.

[3] H. Schwenk; Computer Speech and Language, 21, 492, 2007.

주체105(2016)년 12월 5일 원고접수

# A Method for Constructing Large Scale of Recurrent Neural Network Language Model for Korean Continuous Speech Recognition

*Ri Hyon Sun, Mun Song Il*

We studied the improving method of training speed to construct language model of large vocabulary continuous speech recognition system using recurrent neural network.

We improved the training speed of language model significantly by using recurrent neural network with divided output layer and enabled to use it to training of large scale of language model for Korean continuous speech recognition.

Key words: recurrent neural network, language model, speech recognition