재귀신경망언어모형을 리용한 미지단어확률추정방법

리현순, 김청일

경애하는 최고령도자 **김정은**동지께서는 다음과 같이 말씀하시였다.

《첨단돌파전을 힘있게 벌려야 나라의 과학기술전반을 빨리 발전시키고 지식경제의 토대를 구축해나갈수 있습니다.》

음성인식체계의 언어모형에서 임의의 모든 어휘들을 학습하는것은 불가능하다. 따라서 음성인식언어모형분야에서 미지단어처리는 중요한 문제로 제기된다. 최근 음성인식체계들에서는 지금까지 많이 사용되여오던 재귀신경망을 언어모형학습에 리용하여 그 성능을 개선하고있다.

재귀신경망언어모형(RNNLM: Recurrent Neural Network Language Model)[1, 2]은 언어의 먼거리(Long－Distance)문맥특징정보들을 반영할수 있는 우월한 모형이다. 그러나 재귀신경망언어모형분야에서는 미지단어처리에 대한 연구가 많이 진행되지 못하였다.

재귀신경망언어모형에서는 모든 미지단어들을《unk》기호로 표기하고 하나의 클라스로 취급하였지만 이 방법은 합리적이지 못하다. 그것은 모든 미지단어들이 동등하게 출현한다고 가정하고있기때문이다.

론문에서는 재귀신경망언어모형을 리용하는데서 나서는 새로운 미지단어처리방법을 제안하였다.

1. 미지단어와 류사한 어휘내 단어목록생성

류사목록생성방법은 다음과 같다.

먼저 인식대상으로 되는 코퍼스를 리용하여 미지단어(OOV: Out－Of－Vocabulary)목록을 생성한다.

다음 매 OOV단어들에 대하여 M개의 어휘내(IV: In Vocabulary) 단어목록을 생성한다. 이 단어목록을 류사목록이라고 한다.

류사목록은 다음과 같이 정의된다.

$$\text{류사목록}(OOV) = \{(IV_1, v_1), (IV_2, v_2), \cdots, (IV_M, v_M)\} \quad (1)$$

$$g(OOV, IV_i) = v_i \quad (2)$$

여기서 $g(\cdot)$는 i번째 IV단어와 OOV사이의 류사도함수이다. v_i는 i번째 IV단어와 OOV사이의 류사도값에 대응된다. 같은 OOV고유명사에 대한 모든 류사도값들의 합은 1이다.

론문에서는 N－그람빈도에 기초하여 류사목록을 생성하고 그것의 상대빈도값을 류사도값으로 리용한다.

어휘내에 있는 단어 w가 OOV고유명사와 같은 문맥에서 출현한다면 w는 OOV고유명사에 대한 류사단어로 리용될수 있다.

OOV고유명사에 대한 류사목록을 찾기 위해 k－그람《$w1, \cdots, OOV, \cdots, wk$》에 대응하여 앞선 단어들과 다음에 놓이는 단어들은 같고 가운데의 OOV고유명사가 w로 교체되는 모든 k－그람들《$w1, \cdots, w, \cdots wk$》의 빈도를 구한다. 가장 높은 빈도를 가진 M개의

가운데 단어들이 이 *OOV*고유명사에 대한 류사목록으로 된다. 류사목록은 재귀신경망언어모형을 리용한 *OOV*고유명사확률추정에 리용된다.

2. 재귀신경망언어모형을 리용한 미지단어확률추정

1) 단어리력에 미지단어가 있는 경우 확률 $P(W(t+1)\,|\,OOV,\,h(t-1))$의 추정방법

*OOV*고유명사는 어휘에 없기때문에 재귀신경망언어모형은 그것에 대응하는 입력세포를 가지지 않는다. 매 *OOV*고유명사를 그것의 류사목록에 있는 *IV*단어들을 리용하여 표현한다.

실례로 *OOV*고유명사의 류사목록이 2개의 *IV*단어들을 포함한다고 하면

$$\text{류사목록}(OOV) = \{(IV_1,\,0.6),\,(IV_2,\,0.4)\} \tag{3}$$

와 같이 표시된다. 이때 이 *OOV*고유명사의 재귀신경망입력벡토르는

$$w(t) = (0\cdots 0\;\;0.6\;\;0\cdots 0\;\;0.4\;\;0\cdots 0) \tag{4}$$

과 같이 표시된다. 여기서 0.6과 0.4는 2개의 *IV*단어들의 류사도값과 그것들의 위치에 대응한다. 이 경우 *OOV*는 류사목록의 *IV*단어들의 선형결합으로 볼수 있다. 만일 류사목록이 *M*개 단어들을 포함한다면 *M*개 *IV*단어들이 모두 사용된다. 재귀신경망의 출력층에서는 언어모형확률 $P(w(t+1)\,|\,OOV,\,h(t-1))$이 출력된다.

2) 예측되여야 하는 단어가 미지단어인 경우 확률 $P(OOV\,|\,w_t,\,h(t-1))$의 추정방법

*OOV*고유명사는 어휘에 없기때문에 재귀신경망언어모형은 그것에 대응하는 출력세포를 가지지 않는다. *OOV*의 확률은 류사목록에 있는 *IV*단어들의 확률을 리용하여 표현한다.

*IV*단어에 대하여 그자체와 그것이 류사목록으로 되는 모든 *OOV*고유명사들을 포함하는 클라스를 다음과 같이 정의한다.

$$\begin{aligned}\text{class}(IV) = \{&IV\text{와 그 }IV\text{가 }OOV\text{고유명사들의 류사목록에}\\ &\text{들어있는 모든 }OOV\text{고유명사}\}\end{aligned} \tag{5}$$

실례로 2개의 *OOV*단어 《함흥, 남포》가 있다고 하자. 《함흥》에 대한 류사단어들은 어휘내의 《평양, 평성》이다.

$$\text{류사목록(함흥)} = \{(\text{평양},\;0.6),\,(\text{평성},\;0.4)\} \tag{6}$$

《남포》의 류사단어들은 《평양, 평성, 신의주》이다.

$$\text{류사목록(남포)} = \{(\text{평양},\;0.5),\;(\text{평성},\;0.3),\;(\text{신의주},\;0.2)\} \tag{7}$$

《평양》과 《평성》에 대한 클라스를 다음과 같이 정의한다.

$$\text{class(평양)} = \{\text{평양},\;\text{함흥},\;\text{남포}\} \tag{8}$$

$$\text{class(평성)} = \{\text{평성},\;\text{함흥},\;\text{남포}\} \tag{9}$$

*OOV*고유명사 《함흥》의 확률 $P(\text{함흥}\,|\,w_t,\,h(t-1))$을 다음과 같이 계산할수 있다.

$$\begin{aligned}P(\text{함흥}\,|\,w_t,\,h(t-1)) = {}& P(\text{class(평양)}\,|\,w_t,\,h(t-1))\times P(\text{함흥}\,|\,\text{class(평양)},\,w_t,\,h(t-1)) +\\ &+ P(\text{class(평성)}\,|\,w_t,\,h(t-1))\times P(\text{함흥}\,|\,\text{class(평성)},\,w_t,\,h(t-1))\end{aligned} \tag{10}$$

$P(\text{class(평양)}\,|\,w_t,\,h(t-1))$과 $P(\text{class(평성)}\,|\,w_t,\,h(t-1))$은 재귀신경망언어모형확률이다.

$P(\text{함흥}\,|\,\text{class(평양)},\,w_t,\,h(t-1))$과 $P(\text{함흥}\,|\,\text{class(평성)},\,w_t,\,h(t-1))$을 다음과 같이 계산할수 있다.

$$P(\text{함흥}\,|\,\text{class}(\text{평양}),\, w_t,\, h(t-1)) = (1-\alpha)\times \times g(\text{함흥},\, \text{평양})/(g(\text{함흥},\, \text{평양}) + g(\text{남포},\, \text{평양})) \tag{11}$$

$$P(\text{함흥}\,|\,\text{class}(\text{평성}),\, w_t,\, h(t-1)) = (1-\alpha)\times \times g(\text{함흥},\, \text{평성})/(g(\text{함흥},\, \text{평성}) + g(\text{남포},\, \text{평성})) \tag{12}$$

α는 class(IV)에 속하는 IV에 해당한 확률몫의 비률을 나타낸다. $(1-\alpha)$는 class(IV)에 속하는 OOV에 해당한 확률몫의 비률을 나타내며 이 무게는 실험적으로 결정된다.

class(IV) 당 서로 다른 α값을 가지는것이 리상적이지만 이 파라메터들을 정확히 추정하는것은 어렵다. 이로부터 모든 단어들에 대하여 실험적으로 하나의 α값만을 결정한다. 즉

$$P(\text{평양}\,|\,w_t,\, h(t-1)) = P(\text{class}(\text{평양}))\times\alpha \tag{13}$$

이다. 모든 단어들에 대한 합이 1이 되도록 정규화하면 다음과 같다.

$$\sum_{m\in IV} P(m\,|\,w_t,\, h(t-1)) + \sum_{m\in OOV} P(m\,|\,w_t,\, h(t-1)) = 1 \tag{14}$$

3. 성 능 평 가

언어모형학습자료로서 40M단어들로 구성되는 《로동신문》본문코퍼스(corpus)를 리용하였으며 생성된 어휘크기는 62K이다.

OOV목록생성을 위한 학습자료로 《체육신문》본문코퍼스(10M단어들로 구성)를 리용하여 8 000개의 미지단어를 생성하였다. 최종적으로 확장된 어휘는 70K이다.

평가 및 파라메터결정을 위한 본문자료로 적어도 하나의 미지단어를 포함하는 약 1 000개씩의 문장들을 체육신문에서 선택하였다.

대비를 위한 기준모형들로 선행방법 1(모든 미지단어들을 unk로 취급, 300개의 클라스들과 숨은 층의 크기 500에 대하여 Mikolov가 제안한 도구를 리용하여 통계적그라디엔트하강을 가진 표준역전파알고리듬으로 학습, 어휘모임크기는 62K), 선행방법 2(미지단어들을 어휘에 포함하여 학습, 어휘모임크기는 70K)를 리용하였다.

먼저 최량인 류사목록개수를 결정한다. 류사목록개수를 1부터 30까지 늘이면서 분기수를 평가하였다.

류사단어개수에 따르는 분기수를 그림 1에 보여주었다.

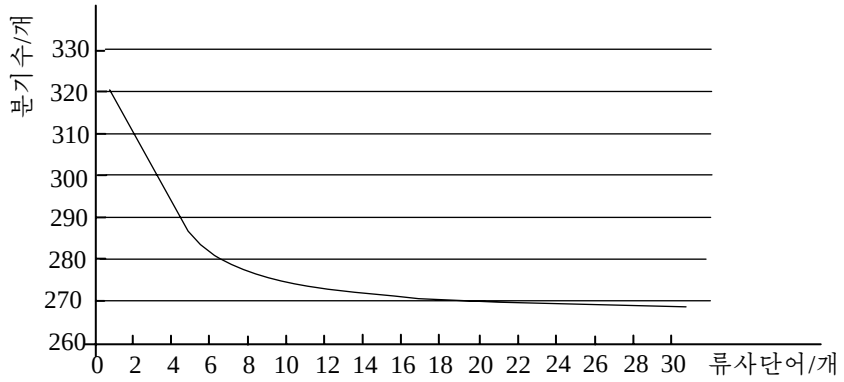

그림 1. 류사단어개수에 따르는 분기수

실험에서는 최량인 류사단어목록의 개수를 26으로 결정하였다.

다음 정규화결수 α를 결정한다. 평가 및 파라메터결정을 위한 본문자료에 대하여 α값을 늘이면서 분기수를 평가하였다.

분기수 대 정규화결수를 그림 2에 보여주었다.

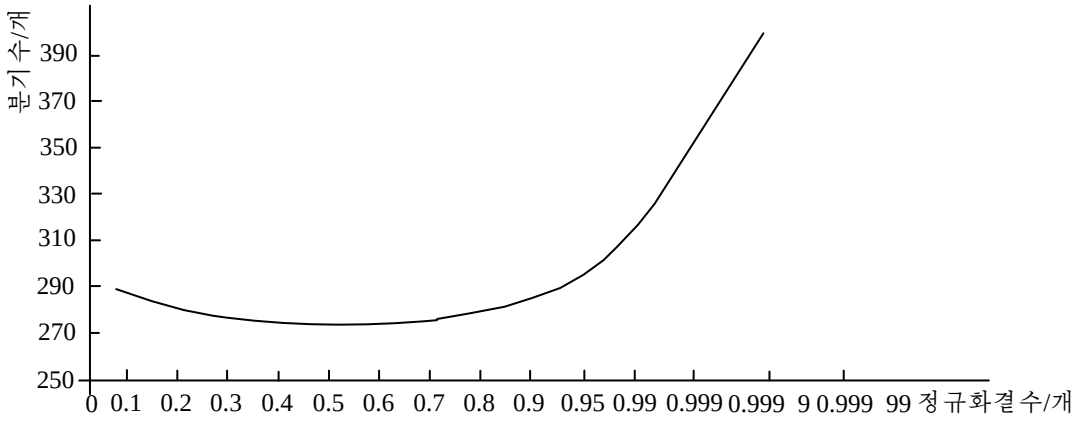

그림 2. 분기수 대 정규화결수

실험에서는 0.6을 α값으로 결정하였다. 이것은 class(IV)의 IV에 부여된 확률이 0.6이라는것을 의미한다.

다음으로 실험에서 결정된 최량파라메터들을 리용하여 제안한 방법과 기준언어모형들사이의 성능을 분기수와 음성인식오유률의 견지에서 평가하였다.

여러 언어모형들의 *OOV*단어확률추정에 대한 성능평가를 다음의 표에 보여주었다.

표. 여러 언어모형들의 *OOV*단어확률추정에 대한 성능평가

모 형	분기수/개	단어오유률/%
선행방법 1	299.5	3.7
선행방법 2	265.2	2.5
제안한 방법	258.6	2.3

실험으로부터 미지단어처리방법이 분기수측면에서 상대적으로 최대 13.6%까지, 단어오유률측면에서 최대 1.4%까지 개선되였다는것을 알수 있다.

맺 는 말

현재 인식과 관련이 있는 새로운 미지단어들을 음성인식체계에 추가하고 그것의 확률분포를 이미 학습된 재귀신경망언어모형을 리용하여 추정하는 언어모형의 동적갱신방법을 제안하였다.

실험을 통하여 제안한 방법이 선행방법보다 미지단어처리에서 효과가 있다는것을 확인하였다.

참 고 문 헌

[1] **김일성**종합대학학보(자연과학), 63, 4, 36, 주체106(2017).

[2] Xunying Liu et al.; IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24, 8, 1438, 2016.

주체108(2019)년 11월 5일 원고접수

The Out-Of-Vocabulary Words Probability Estimation by Using Recurrent Neural Network Language Model

Ri Hyon Sun, Kim Chong Il

In this paper, we propose a new probability estimating method of out-of-vocabulary words by using recurrent neural network language model.

Keywords: Language modeling, speech recognition, OOV language modeling