2중기능조건부우연마당을 리용한 고유실체인식정확도 개선의 한가지 방법

우 준 민

고유실체인식에 많이 리용되는 통계모형들[3]로는 숨은마르꼬브모형(HMM), 최대엔트로피마르꼬브모형(MEMM), 조건부우연마당(CRF) 등이 있다.

선행연구[1]에서는 조선어고유실체인식정확도를 개선할수 있게 실체지시단어모임을 리용한 특징들을 추가하고 CRF를 리용하여 인식을 진행하였다.

선행연구[2]에서는 클라스*n*그람언어모형을 구축하고 그것에 의한 어휘분리방법으로 형태부해석과 고유실체인식의 일체화를 실현하는 방법으로 조선어고유실체인식을 진행하였다. 이 방법은 형태부해석과 고유실체인식을 동시에 진행한다는 우점을 가지고있지만 고유실체인식의 고유한 언어적정보들을 모형에로 부호화하기가 힘들고 훈련본문에서 매우 적은 빈도수를 가지고 나타나거나 전혀 나타나지 않는 고유실체의 특성을 원만히 고려하지 못하는 결함을 가지고있다.

이러한 문제를 해결하기 위하여 다음과 같은 문제를 설정한다.

첫째로, 조선어고유실체의 문맥특성을 고려한 2중기능CRF모형을 제기한다.

둘째로, 2중기능CRF모형의 학습과 추론방법을 제기하고 모형을 리용한 고유실체인식방법을 제기한다.

셋째로, 실험을 통하여 제안한 방법의 효과성을 검증한다.

1. 2중기능조건부우연마당을 리용한 조선어고유실체인식방법

관측렬 $x=([x_t^{\mathrm{T}}]_{t=1}^{Tx})^{\mathrm{T}}$ 이 주어졌을 때 전체 렬에 대한 클라스값 $h\in H$ 와 요구되는 표식렬 $\boldsymbol{y}=[y_t]_{t=1}^{T}$ 들을 동시에 얻는 문제를 생각하자.

이 문제를 해결하기 위하여 다음과 같은 형식의 CRF에 기초한 모형을 설정한다.(그림)

$$p(h,\boldsymbol{y}\mid\boldsymbol{x})=\frac{1}{Z(\boldsymbol{x})}\exp\left[\sum_{t=2}^{T}\phi_t(h,y_t,y_{t-1},\boldsymbol{x}_t)+\phi_1(h,y_1,\boldsymbol{x}_1)\right]$$

여기서

$$Z(\boldsymbol{x})=\sum_{h}\sum_{\boldsymbol{y}}\exp\left[\sum_{t=2}^{T}\phi_t(h,y_t,y_{t-1},\boldsymbol{x}_t)+\phi_1(h,y_1,\boldsymbol{x}_1)\right]$$

이다.

론문에서는 다음과 같은 형식의 클라스조건부포텐샬함수들의 모임을 도입한다.

$$\phi_t(h,y_t,y_{t-1},\boldsymbol{x}_t)=\phi_t^1(h,y_t,\boldsymbol{x}_t)+\phi_t^2(h,y_t,y_{t-1})$$

$$\phi_1(h,y_1,\boldsymbol{x}_1)=\phi_1^1(h,y_1,\boldsymbol{x}_1)+\phi_1^2(h,y_1)$$

여기서 $\varphi_t^2(h, y_t, y_{t-1})$과 $\varphi_t^1(h, y_t, x_t)$는 클라스조건부적인 전이포텐샬함수와 한변수포텐샬함수이다.

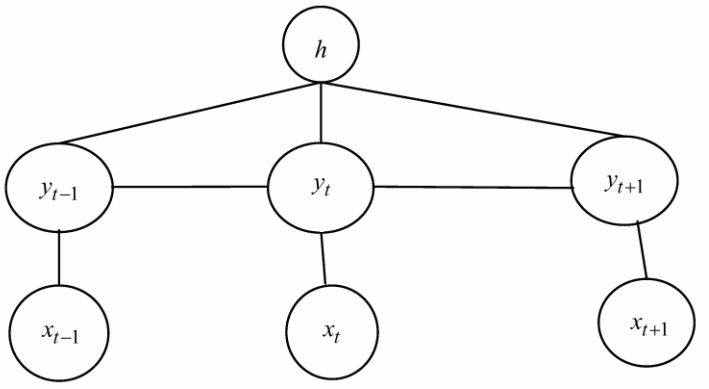

그림. 2중기능조건부우연마당모형

선형제한을 주면 다음과 같이 쓸수 있다.

$$\phi_t^1(h,\ y_t,\ \boldsymbol{x}_t) = \sum_{i=1}^{K} \delta(y_t - i)\boldsymbol{\omega}_i^h \cdot \boldsymbol{x}_t$$

$$\phi_t^2(h, y_t, y_{t-1}) = \sum_{i=1}^{K}\sum_{j=1}^{K} \delta(y_t - j)\delta(y_{t-1} - i)\xi_{ij}^h$$

$$\phi_1^2(h, y_1) = \sum_{i=1}^{K} \delta(y_t - j)\xi_i^h$$

$$\phi_1(h, y_1, \boldsymbol{x}_1) = \phi_1^1(h, y_1, \boldsymbol{x}_1) + \phi_1^2(h, y_1)$$

모형학습은 모형포텐샬함수들의 파라메터들에 대한 추정으로 이루어진다.

클라스조건포텐샬함수를 고찰할 때 이것은 다음과 같은 파라메터들의 추정으로 된다.

$$\{\boldsymbol{\omega}_i^h\}_{h\in\boldsymbol{H},\ i\in\boldsymbol{y}},\ \{\xi_{ij}^h\}_{h\in\boldsymbol{H},\ i,\ j\in\boldsymbol{y}},\ \{\xi_i^h\}_{h\in\boldsymbol{H},\ i\in\boldsymbol{y}}$$

모형학습을 위하여 다음과 같은 로그우도최량화문제를 풀어보자.

$$\log p(h,\ \boldsymbol{y}\,|\,\boldsymbol{x}) = \sum_{t=2}^{T} \phi_t(h, y_t, y_{t-1}, \boldsymbol{x}_t) + \phi_1(h, y_1, \boldsymbol{x}_1) - \log Z(\boldsymbol{x})$$

여기서 모형의 로그우도를 계산하는것은 y와 h의 전체 값에 대하여 합인 $Z(x)$에 대한 계산을 요구한다. 이 계산을 쉽게 하기 위하여 다음과 같은 형식으로 $Z(x)$를 표현한다.

$$Z(\boldsymbol{x}) = \sum_{h\in H} Z(\boldsymbol{x}\,|\,h)$$

여기서

$$Z(\boldsymbol{x}\,|\,h) = \sum_{\boldsymbol{y}} \exp\left[\sum_{t=2}^{T} \phi_t(h, y_t, y_{t-1}, \boldsymbol{x}_t) + \phi_1(h, y_1, \boldsymbol{x}_1)\right]$$

이다.

이것은 앞방향－뒤방향알고리듬에 의하여 쉽게 계산할수 있다.

$$Z(\boldsymbol{x}\,|\,h)=\sum_{j=1}^{K}\alpha_t(h,\ j)$$

웃식에서 클라스조건부앞방향확률 $\alpha_t(h,\ j)$는 다음과 같이 계산된다.

$$\alpha_t(h,\ j)=\sum_{j=1}^{K}\alpha_{t-1}(h,\ j)\exp\{\phi_t(h,\ y_t=j,\ y_{t-1}=i,\ \boldsymbol{x}_t)\},\ t\geq 2$$

초기값은

$$\alpha_1(h,\ j)=\exp\{\phi_1(h,\ y_1=j,\ \boldsymbol{x}_1)\}$$

이다.

*N*개의 훈련렬에 대하여 모형의 로그우도를 최대화하는 파라메터들은 L-BFGS와 같은 방법으로 구할수 있다.

렬에 대한 분류 다시말하여 관측렬 $\boldsymbol{x}$가 주어졌을 때 최량클라스표식 h는 다음과 같은 최량화문제를 풀어 구한다.

$$\hat{h}=\arg\max_{h\in H}p(h\,|\,\boldsymbol{x})$$

여기서

$$p(h\,|\,\boldsymbol{x})=\sum_{\boldsymbol{y}}p(h,\ \boldsymbol{y}\,|\,\boldsymbol{x})=\frac{\sum_{\boldsymbol{y}}\exp\left[\sum_{t=2}^{T}\phi_t(h,\ y_t,\ y_{t-1},\ \boldsymbol{x}_t)+\phi_1(h,\ y_1,\ \boldsymbol{x}_1)\right]}{Z(\boldsymbol{x})}=\frac{Z(\boldsymbol{x}\,|\,h)}{Z(\boldsymbol{x})}$$

이다.

따라서 이 문제는 다음과 같이 쓸수 있다.

$$\hat{h}=\arg\max_{h\in H}Z(\boldsymbol{x}\,|\,h)$$

관측렬 $\boldsymbol{x}$가 주어졌을 때 렬토막화는 다음과 같은 최량화문제로 구성되는 렬복호화문제로 볼수 있다.

$$\hat{y}=\arg\max_{\boldsymbol{y}}\log p(\boldsymbol{y}\,|\,\boldsymbol{x})$$

이렇게 구성한 2중기능조건부우연마당을 리용하면 고유실체와 형태부들이 결합하여 생기는 고유실체들을 인식할수 있다.

이때 x는 단어렬이고 y는 형태부렬에 대한 표기이며 h는 그 형태부렬에 대한 인식된 고유실체인식표식이다.

다음으로 고유실체인식을 위하여 조건부우연마당에서 리용하는 징표형타선택문제를 생각하자.

우선 관측렬에 대한 징표선택에서 가장 중요한것은 이전의 표식이다.

실례로《황해북도 사리원시 ×× 협동농장》이라고 하면《×× 협동농장》이 조직실체인것으로 해서 전체적으로 조직실체지시단어가 된다.

또한 중요한 징표의 하나는 의미정보이다.

실례로《황해남도 옹진군 ×× 수산사업소》라고 하면《×× 수산사업소》의 의미범주가 조직에 속하기때문에 전체적으로 조직실체지시단어가 된다.

조선어고유실체인식에 리용하는 징표형태는 표 1과 같다.

표 1. 조선어고유실체인식에 리용하는 징표형태

징표형태	징표의 의미
CUR_ORG	현재 단어가 조직실체이고 렬의 마지막단어인 경우
CUR_MEAN_ORG	현재의 단어의 의미범주가 조직의 하위개념이고 렬의 마지막단어인 경우
PREV_ORG	이전단어의 표식과 현재단어의 표식
PREV_MEAN_ORG	이전단어의 의미와 현재단어의 의미
PREV_TAG_MEAN_ORG	이전단어의 표식과 현재단어의 의미

이와 같은 형태의 징표를 장소와 사람범주에 대하여 선택하고 2중기능CRF를 리용하여 단어렬에 대하여 고유실체인식을 진행하면 우에서 제기한 문제가 해결될수 있다.

2. 실험 및 성능평가

숨은마르꼬브모형과 규칙에 기초한 고유실체인식기와의 인식성능비교를 진행한 결과는 표 2와 같다. 실험에서는 확장된 고유실체지시단어모임을 리용하여 고유실체인식을 진행하였다.

표 2. 인식성능비교

구분	적중률	완전률	*F*-척도
HMM	88.78	82.35	85.44
선행방법[2]	98.9	69	83.86
제안방법	93.75	98.14	91.39

표 2에서 보는바와 같이 추출규칙에 기초한 방법은 적중률은 높지만 완전률이 낮다는것을 알수 있다. 그것은 추출규칙에 반영하지 못한 고유실체들이 존재하기때문이다. 한편 최대엔트로피모형에 기초한 방법은 숨은마르꼬브모형에 기초한 방법보다 적중률과 완전률을 둘 다 개선하였다. 론문에서 제기한 방법은 또한 추출규칙에 기초한 방법에 비하여 적중률은 약간 떨어지지만 완전률을 훨씬 높였으며 전체적인 체계의 성능평가지표인 *F*-척도를 개선하였다.

맺 는 말

조선어문맥특성을 반영하여 고유실체인식정확도를 개선할수 있는 2중기능조건부우연마당을 설계하고 그것을 리용한 고유실체인식방법을 제기하고 효과성을 검증하였다.

참 고 문 헌

[1] **김일성**종합대학창립70돐기념 전국부문별과학기술토론회론문집(정보, 자동화), 31, 주체105(2016).

[2] 홍충성; 정보과학, 4, 45, 주체104(2015).

[3] D. Kaur, V. Gupta; International Journal of Computer Science Issues, 7, 239, 2010.

주체107(2018)년 11월 5일 원고접수

A Method of Improving the Accuracy of the Named Entity Recognition Using Dual-functional CRF

U Jun Min

In this paper we designed the dual-functional CRF by taking into account Korean named entity characteristics and proposed a method of improving the performance of NER systems using it.

Key words: named entity recognition, conditional random field