ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2018 March: 24(1): 253~267 http://dx.doi.org/10.13088/jiis.2018.24.1.253

온톨로지 지식 기반 특성치를 활용한 Bidirectional LSTM-CRF 모델의 시퀀스 태깅 성능 향상에 관한 연구*

진승희
연세대학교 정보산업공학과
(*seunghe216@gmail.com*)

장희원
연세대학교 정보산업공학과
(*hwjang1000@gmail.com*)

김우주
연세대학교 정보산업공학과
(*wkim@yonsei.ac.kr*)

본 연구는 질의 응답(QA) 시스템에서 사용하는 개체명 인식(NER)의 성능을 향상시키기 위하여 시퀀스 태깅 방법론을 적용한 새로운 방법론을 제안한다. 사용자의 질의를 입력 받아 데이터베이스에 저장된 정답을 추출하기 위해서는 사람의 언어를 컴퓨터가 알아들을 수 있도록 구조화 질의어(SQL)와 같은 데이터베이스의 언어로 전환하는 과정이 필요한데, 개체명 인식은 사용자의 질의에서 데이터베이스에 포함된 클래스나 데이터 명을 식별하는 과정이다. 기존의 데이터베이스에서 질의에 포함된 단어를 검색하여 개체명을 인식하는 방식은 동음이의어와 문장성분 구를 문맥을 고려하여 식별하지 못한다. 다수의 검색 결과가 존재하면 그들 모두를 결과로 반환하기 때문에 질의에 대한 해석이 여러 가지가 나올 수 있고, 계산을 위한 시간복잡도가 커진다. 본 연구에서는 이러한 단점을 극복하기 위해 신경망 기반의 방법론을 사용하여 질의가 가지는 문맥적 의미를 반영함으로써 이러한 문제를 해결하고자 했고 신경망 기반의 방법론의 문제점인 학습되지 않은 단어에 대해서도 문맥을 통해 식별을 하고자 하였다. Sequence Tagging 분야에서 최신 기술인 Bidirectional LSTM-CRF 모델을 도입함으로써 신경망 모델이 가진 단점을 해결하였고, 학습되지 않은 단어에 대해서는 온톨로지 기반 특성치를 활용하여 문맥을 반영한 추론을 사용하였다. 음악 도메인의 온톨로지(Ontology) 지식베이스를 대상으로 실험을 진행하고 그 성능을 평가하였다. 본 연구에서 제안한 방법론인 L-Bidirectional LSTM-CRF의 성능을 정확하게 평가하기 위하여 학습에 포함된 단어들뿐만 아니라 학습에 포함되지 않은 단어들도 포함한 질의를 평가에 사용하였다. 그 결과 L-Bidirectional LSTM-CRF 모형을 재학습 시키지 않아도 학습에 포함되지 않은 단어를 포함한 질의에 대한 개체명 인식이 가능함을 확인하였고, 전체적으로 개체명 인식의 성능이 향상됨을 확인할 수 있었다.

주제어 : 시퀀스 태깅, CRF(Conditional Random Field), LSTM(Long Short Term Memory), 질의응답 시스템, 온톨로지

논문접수일 : 2017년 11월 13일 논문수정일 : 2018년 3월 20일 게재확정일 : 2018년 3월 23일
원고유형 : 일반논문 교신저자 : 김우주

1. 서론

1.1 연구 개요

음성인식 기반 QA시스템은 사용자의 질의를 입력 받아 데이터베이스에서 정답을 찾아 준다. 여기서 사용자의 질의가 자연어이기 때문에 데이터베이스가 이해할 수 있는 SQL과 같은 명령어로 변환하는 과정이 필요하다. 예를 들어 '소

* 본 연구는 SK Telecom과 'LTE 장비 및 IT 인프라 자산의 효율적 운영을 위한 자가진단 통합 제어 관제 시스템' 사업의 지원을 받았습니다.

녀시대가 부른 노래는?’이라는 질의는 최소 ‘Select * from 노래 where 가수 = “소녀시대”’와 같은 명령어로 전환이 되어야 한다. 이를 위해서 사용자 질의에서 데이터베이스와 매칭시킬 수 있는 요소를 찾아내는 것이 중요하다.

지식기반의 데이터베이스인 Ontology에서 질의 내 단어들을 인스턴스로 갖는 클래스 정보로 개체명를 인식할 것이다. 위의 예시에서 보면 “소녀시대”가 가수로 인식되었다. 하지만 Ontology에는 “소녀시대”가 곡명이란 정보도 들어있는 경우가 있다. 이렇게 되면 “소녀시대”가 가수로 인식되어야 하는지 곡으로 인식되어야 하는지 판단이 어렵다. 따라서 기존에 사용하던 인덱스를 기반의 검색으로 개체명을 인식을 하는 경우에 가수와 곡이라는 검색 결과가 존재하면 두 가지를 결과로 반환하기 때문에 질의에 대한 해석이 여러 가지가 나올 수 있고, 계산을 위한 시간복잡도가 커진다. 기존 QA시스템에서 청킹은 사용자 질의로부터 나올 수 있는 모든 경우의 부분집합을 구한 후 검색되는 문장성분 구 중에 가장 긴 구가 선택될 확률이 높았다. 이러한 청킹 처리는 알고리즘도 복잡하지만 질의에 따라 항상 정답인 것은 아니라는 문제점이 존재한다. 따라서 위의 문제들을 해결하고자 신경망 모델 연구를 알아보고, 시퀀스 태깅 중 최신 기술 방법론을 적용해 본다.

1.2 연구 내용

먼저 동음이의어 문제는 신경망 기반의 Bidirectional LSTM-CRF을 적용하면 문맥에 따른 개체명 인식이 가능하다(Graves et al., 2013). 예를 들면 두 개의 질의 ‘소녀시대가 부른 곡 장르는?’, ‘소녀시대 작곡가가 누구야?’가 있을 때 ‘소녀시대’라는 개체를 가수 이름으로 인식할 것인지 노래 이름으로 인식할 것인지를 판단해 준다. 또한, 청킹 문제도 효율적인 방법으로 결과를 낼 수 있다. 예를 들면 질의 ‘널 부르다를 부른 가수는’에 대해 형태소 분석을 하게 되면 ‘널 부르다 부르다 가수’가 되는데 이 경우 앞의 ‘부르다’는 곡명으로 문장성분 구에 속해야 하지만 뒤의 ‘부르다’는 서술어로 개체명 인식이 되지 않아야 하는 경우가 있다. 신경망 모델은 이 부분을 해결해 준다.

위의 신경망 모델을 적용한 이유는 다음과 같다. 문장 내에서 Tag 시퀀스정보를 고려하기 위해 Chain-structured Conditional Random Field(CRF) 층을 사용하였다(Lafferty et al., 2001). 또한, 전통적인 Recurrent neural networks(RNN)보다 중요한 문맥적 의미를 오래 기억할 수 있는 동시에 과거와 미래의 요소들을 모두 고려할 수 있도록 Bidirectional Long Short Term Memory(BI-LSTM)을 결합하였다(Elman, 1990; Graves et al., 2013). 추가적으로 본 논문에서는 학습되지 않은 단어에 대해서는 온톨로지 기반 특성치를 활용하여 신경망 기반 방법론의 문제점을 해결하고자 하였다. 예를 들어 학습되지 않았던 ‘이미자’라는 새로운 단어가 포함된 질의 ‘이미자가 부른 곡 장르는?’에 대해서 ‘이미자’개체가 가수 이름으로 정확히 판별하는 것을 목표로 한다.

실험 데이터는 SKT Music ontology를 사용했다. 식별 Tag는 MusicArtist, Track, Genre, MusicGroup, MusicAlbum, Person, MusicActivity, Country, Instrumental, Organization으로 총 10개를 사용하였고, 식별되지 않는 것은 Other(O)로 표현하였다. 또한, 문장성분 구를 위해 Tag의 첫 번째 Tag에는 B-를 추가하고, 이 후 Tag에는 I-를 추가하여 청킹을 처리하였다. 실험 결과는 이러

한 지식기반의 질의에 대해 기존 인덱스를 기반으로 개체명 인식을 하던 방법보다 Bidirectional LSTM-CRF모델이 문맥적 의미를 반영하여 나은 성능을 보였다. 또한, 학습되지 않은 단어가 포함된 질의에 대해서는 새롭게 제안된 모델을 통해 문맥적 의미를 반영한 추론이 가능했다.

2. 관련연구

2장에서는 신경망 모델을 이용한 개체명 인식과 관련된 연구를 살펴 본다. 또한, CRF모델과 Bidirectional LSTM모델, 그리고 두 모델을 결합한 Bidirectional LSTM-CRF모델을 살펴 본다.

2.1 시퀀스 태깅

개체명 인식은 자연어를 처리하는 분야이다. 자연어 처리의 기본 과정은 사람의 언어를 기계가 이해할 수 있는 형태로 먼저 전환하는 것인데, 이를 임베딩이라고 한다. 언어를 임베딩하는 과정은 크게 세가지로, 서로 다른 단어를 임의의 숫자의 조합으로 표현하는 방식과 문자열을 인식해서 식별하는 방법, 뜻을 가지는 단어 단위로 식별하는 방법이 있다. 문자열을 인식해서 단어를 식별하는 모델은 딕셔너리를 사용하지 않을 수 있고 단어를 임베딩하는 것에 비해 의미 보다는 형태상 유사성을 잘 반영한다(Zhang et al.. 2015; Kim et al.. 2016). Word2vec으로 대표되는 단어를 임베딩하는 모델은 단어의 모양이 아닌 문맥 속에서 단어의 의미를 벡터로 표현한다(Mikolov et al.. 2013). 많은 개체명 인식 연구에서 전처리로 임베딩 기술을 사용하는 것이 일반적이지만 일부 LSTM기반 모델에서는 실효성이 없다고 한다(Huang et al.. 2015).

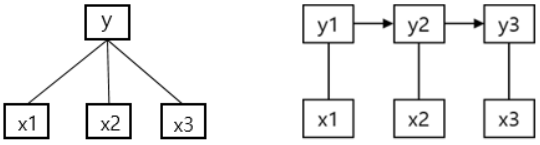
LSTM은 RNN과 같은 시퀀스 모델의 장기 기억 문제를 어느 정도 해결한 모델로 시퀀스 데이터를 다루는데 많이 쓰인다. 개체명 인식 분야에서도 단어 자체만이 아니고 문맥을 반영하기 위해 많이 사용되는 모델이다. LSTM은 긴 시퀀스를 처리하지만 최근 입력에 편향되는 경향을 보이기도 한다(Lample et al.. 2016). 따라서 이전의 정보뿐만 아니라 양방향으로 정보를 반영할 수 있는 bi-directional LSTM을 많이 활용한다. 여기서 그치지 않고 bi-directional LSTM 모델에 CNN 기반 임베딩을 활용해 개체명 인식의 성능을 향상 시키려는 시도가 있었다(Chiu et al.. 2015). 또한 CRF와 bi-directional LSTM을 결합하여 임베딩 기술에 의존하지 않고 기존의 bi-directional LSTM보다 더 좋은 성능을 보이는 연구도 있었다(Huang et al.. 2015).

2.2 Bidirectional LSTM-CRF

2.2.1 Chain-structured Conditional Random Field(CRF)

독립적인 feature set X로부터 target 클래스 Y를 예측해야 한다고 할 때, 결합확률을 구하는 대신 정규화 트릭을 사용하면 X가 주어졌을 때 Y가 나올 확률 값 P(Y|X)을 구할 수 있다. 각 x와 Y와의 관계를 곱하게 되면 정규화되지 않은 확률 $\tilde{P}(\mathrm{Y},\mathrm{X})$를 구할 수 있고, X가 독립이라는 가정 때문에 X에 대한 부분확률 $\tilde{P}(\mathrm{X}) = \sum_{Y}\tilde{P}(\mathrm{Y},\mathrm{X})$을 구할 수 있다. 따라서 정규화를 하면 구하려고 했던 확률 값 $\mathrm{P}(\mathrm{Y}|\mathrm{X}) = \tilde{P}(\mathrm{Y},\mathrm{X})/\tilde{P}(\mathrm{X})$을 구할 수 있게 된다. <Figure 1> 에서 왼쪽은 단일 y일 때의 CRF 네트워크를 보여주며 오른쪽은

sequence y에 대해 Chain-structured CRF 네트워크를 보여준다.



〈Figure 1〉 CRF and Linear-chain CRF

그럼 Y가 연속적인 경우를 살펴보자. $X = (x_1, x_2, \cdots, x_n)$, $y = (y_1, y_2, \cdots, y_n)$라고 할 때, X가 독립이기 때문에 y_2는 x_2로부터 그리고 y_1으로부터 영향을 받게 된다. 즉 x로부터 y와의 관계를 하나의 함수로 정의할 수 있고, 이전 시간대의 y와 현재의 y와의 관계를 또 하나의 함수로 정의할 수 있다. 다음 식(1)은 정규화되지 않은 확률을 구하기 위해 각 함수들을 곱한 것이다. 스코어 S값은 지수화하여 사용될 것이기 때문에 각 항들을 덧셈으로 표현할 수 있다.

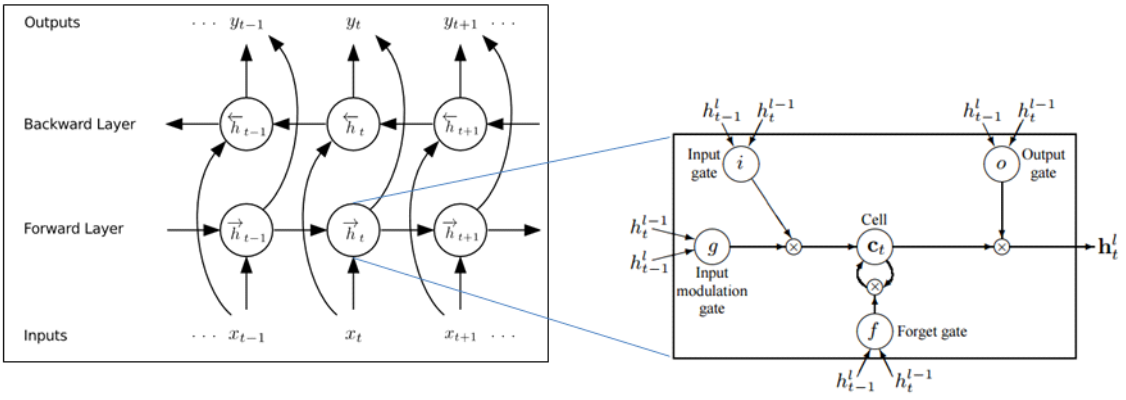
$$S(X, y) = \sum_{i=0}^{n} T_{y_i, y_{i-1}} + \sum_{i=1}^{n} E_{i, y_i} \quad (1)$$

n은 각 문장의 단어 수를 의미한다. k를 가능한 Tag의 숫자라고 할 때, **T**의 차원은 k * k와 같고, y_{i-1}에서 y_i Tag가 나올 전이 값들이 채워진다. E의 차원은 k * n으로 i번째 x로부터 y_i가 나올 예측 값 즉, Bidirectional LSTM output을 의미한다(Graves et al., 2013). Y_X를 문장 X로부터 나올 수 있는 모든 경우의 Tag 시퀀스들이라고 한다면 다음과 같은 정규화 식(2)을 통해 P(y|X)를 구할 수 있다.

$$P(y|X) = e^{s(X,y)} \; / \; \sum_{\tilde{y} \in Y_X} e^{s(X,\tilde{y})} \quad (2)$$

2.2.2 Bidirectional LSTM

전통적인 Recurrent Neural Network(RNN)의 장기 의존성 문제를 극복한 모델로 오랫동안 정보를 기억할 수 있다(Elman, 1990). RNN의 cell에 Long Short-Term Memory(LSTM) cell을 적용한 것이다(Elman, 1990; Hochreiter and Schmidhuber, 1997; Graves et al., 2005). 다음 <Figure 2>는 하나의 LSTM cell에 대해 보여준다(Graves et al., 2005).

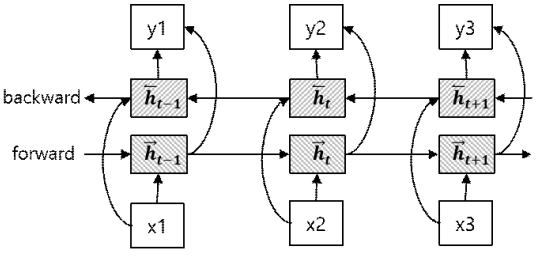


〈Figure 2〉 A Long Short-Term Memory Cell

LSTM memory cell은 다음과 같이 구현된다.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (3)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (4)$$
$$c_t = f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (5)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (6)$$
$$h_t = o_t \tanh(c_t) \quad (7)$$

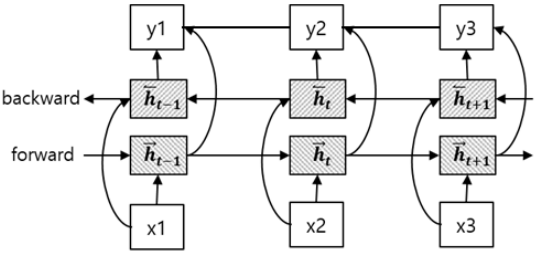
σ는 logistic sigmoid 함수이다. ***i***, ***f***, ***o*** 그리고 ***c***는 각각 input gate, forget gate, output gate 그리고 cell vector들이다. 이 벡터들은 모두 hidden vector ***h***와 크기가 같다. input gate, forget gate를 사용하여 cell vector라는 기억을 저장하는 자료구조에 과거의 정보를 얼마만큼 남겨 둘지 새로운 정보를 얼마만큼 기억해야 정확한 추론을 할 수 있는지를 결정하게 된다. 이러한 특성으로 RNN을 을 비롯한 다른 신경망 모델보다 상대적으로 시퀀스 정보를 잘 다룰 수 있다. 과거와 더불어 미래의 정보를 사용하여 현재를 추론하기 위해 Forward LSTM과 Backward LSTM의 쌍을 엮은 모델을 Bidirectional LSTM이라고 한다(Graves et al., 2013). 과거와 미래의 정보를 사용하여 단어의 Tag를 예측할 수 있을 뿐만 아니라, 문장 내의 Tag 시퀀스 정보도 고려할 수 있다. Bidirectional LSTM의 구조는 다음의 <Figure 3>과 같다.



〈Figure 3〉 A Bidirectional LSTM

2.2.3 Bidirectional LSTM-CRF

2.2.1절에서 소개된 Chain-structured CRF에서 x와 y와의 관계에 대한 함수를 2.2.2절에서 소개된 Bidirectional LSTM으로 정의한 것이 바로 Bidirectional LSTM-CRF이다(Lafferty et al., 2001; Graves et al., 2013; Ling et al., 2015). 다음 <Figure 4>는 Bidirectional LSTM-CRF에 대해 보여준다(Ling et al., 2015).



〈Figure 4〉 A Bidirectional LSTM-CRF

X는 독립이라는 가정이 있기 때문에 연결되지 않은 것을 확인할 수 있다. Backward LSTM output과 forward LSTM output은 CRF층의 input이 되는 것이다(Lafferty et al., 2001). 위 네트워크에 대해 스코어 S값을 정의하자면 다음 식(8)과 같다.

$$s(X, \tilde{y}) = \sum_{i=1}^{n}([T]_{[\tilde{y}]_{i-1},[\tilde{y}]_i} + [f_\theta]_{[\tilde{y}]_i, i}) \quad (8)$$

다음 <Figure 5>는 T행렬의 예시이며, START_TAG와 STOP_TAG를 포함하여 (k+2) * (k+2)의 차원으로 구현된다.

O B-MusicGroup start_tag stop_tag

i \ i-1	0	1	...	k	k+1
0	-0.3826	-0.2655		-0.1781	1.043
1	1.1465	-1.3879		1.1968	1.4066
...					
k	1.4275	1.7098		-0.5496	-0.7545
k+1	1.1578	-0.5768		-0.1954	0.4206

〈Figure 5〉 A Transition Matrix

f_θ는 i번째의 x에서 i번째의 y를 예측할 Bidirectional LSTM의 output값을 의미하며 마찬가지로 START_TAG와 STOP_TAG를 포함하여 n * (k+2)의 차원으로 표현된다(Ling et al., 2015). 다음 <Figure 6>은 '소녀시대'단어에 대한 Bidirectional LSTM의 output 예시이다.

O B-MusicGroup start_tag stop_tag

0	1	...	k	k+1
-0.3826	-0.2654		-1.185	-10000

〈Figure 6〉 '소녀시대' Bidirectional-LSTM output 예시

3. 향상된 Bidirectional LSTM-CRF 방법론

학습시키지 않은 단어가 들어올 때를 대비해 온톨로지 기반 특성치로 온톨로지 지식베이스로부터 획득한 개체명 인식 Tag 정보를 Lucene index DB에 저장하였고, 이를 활용한 신경망 모델을 본 연구에서는 새롭게 제안하였다.

3.1 L-Bidirectional LSTM-CRF

온톨로지에는 새로운 데이터들이 추가될 경우가 생기고, 이 경우에 새롭게 추론 과정을 거쳐야 한다. 다시 학습시켜야 하는 과정을 줄이기 위해 본 연구에서는 방법을 제안한다. 온톨로지는 루씬(Lucene)이라는 검색엔진에 검색을 함으로써 색인을 얻을 수 있다. 즉, 온톨로지에 들어있는 단어들이 어떤 클래스에 속하는지 알 수가 있다. 이 때 클래스는 Tag로 표현할 수 있다. x_i에 학습되지 않았던 새로운 단어가 들어왔을 때 본 연구에서는 루씬으로 얻은 인덱스 데이터베이스를 사용하기로 하였다. 루씬 인덱스 데이터베이스에 있는 Tag 정보를 feature로 학습시키는 파라미터로 사용하였다. X를 루씬 인덱스 데이터베이스에 검색하여 Tag 정보를 찾고, 학습된 Tag feature를 사용하는 것이다. 루씬 인덱스 데이터베이스를 통해 얻은 Tag feature를 L이라고 하고 크기는 k * n과 같다. 스코어 S값을 정의하면 다음 식(9)과 같다.

$$s(X,\tilde{y}) = \sum_{i=1}^{n}\left([T]_{[\tilde{y}]_{i-1},[\tilde{y}]_i} + [f_\theta]_{[\tilde{y}]_i,i} + [L]_{[\tilde{y}]_i,i}\right) \quad (9)$$

Lucene index DB의 예시는 다음 <Figure 7>와 같다. 온톨로지의 인스턴스 정보는 Word로 표현되고, 인스턴스가 속하는 클래스 정보들은 Tag를 인덱스로 표현하여 저장된다.

Word	Tag
지드래곤	5,6
태연	5,6
...	
소녀시대	1,3,4
뱅뱅뱅	3,9
이미자	**4**
재즈	2
서태지와	3
아이들	13

〈Figure 7〉 A Lucene index DB 예시

'이미자'를 예로 들면 아래 <Figure 8>에 나와 있는 것처럼 Tag feature 중 '이미자'에 해당하는 Tag 자리의 정보를 사용한다.

그림으로 위 네트워크를 표현하면 다음 <Figure 9>와 같다. 시간대마다 단어에 해당하는 Tag feature의 정보가 추가로 CRF층에 input으로 들어가게 된다(Lafferty et al., 2001).

4. 실험

4.1 데이터

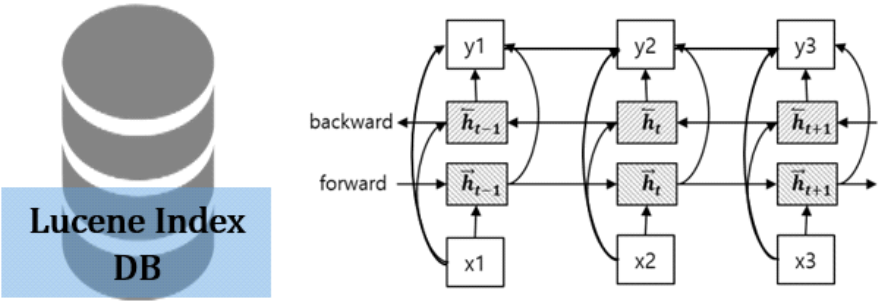
데이터는 SKT 음악 온톨로지를 사용하였다. 총 사용자 질의 문장 수와 토큰 수는 다음 <Table 1>과 같다.

〈Table 1〉 Data set

	Training Set	Test Set
Sentence	910	370
Token	3613	1433

O	B-MusicGroup		B-Track	B-MusicArtist	B-Person		start_tag	stop_tag
0	1	...	3	**4**	5	...	n	n+1
33.534	103.55		46.784	38.918	12.596		-8540.8	-8540.8

〈Figure 8〉 '이미자'의 Tag feature 사용 예시



〈Figure 9〉 A Bidirectional LSTM-CRF model with Lucene Tag Feature

식별 Tag는 MusicArtist, Track, Genre, MusicGroup, MusicAlbum, Person, MusicActivity, Country, Instrumental, Organization으로 총 10개를 사용하였고, 식별되지 않는 것은 Other(O)로 표현하였다. 또한, 문장성분 구를 위해 Tag의 첫 번째 Tag에는 B-를 추가하고, 이 후 Tag에는 I-를 추가하여 청킹을 위한 Tag를 재명명하였다.

4.2 학습

실험 환경은 리눅스(linux)에서 파이선(python) 언어로 Pytorch 라이브러리를 사용하였다. 변수 설정은 Learning rate : 0.01, Hidden layer size : 3, Epoch : 200으로 하였다. 학습은 결국 식(2)를 최대화 시키기 위해 다음의 식(10)을 최소화하는 Stochastic Gradient Descent(SGD)를 하였다.

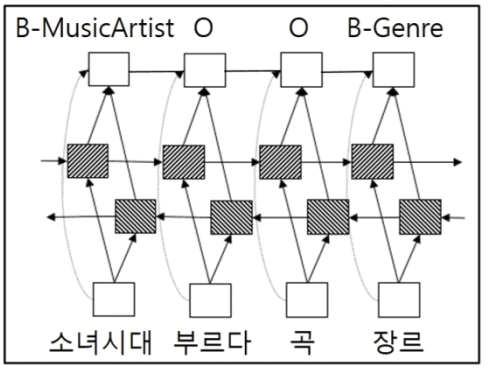
$$- \log(p(y|X)) = \log(\sum_{\tilde{y}\in Y_X} e^{s(X,\tilde{y})}) - s(X,y) \quad (10)$$

스코어 S값을 구할 때 사용되는 T와 f_θ, 그리고 L은 log분포를 따르며 이를 통해 학습되는 파라미터는 T인 전이행렬과 Bidirectional LSTM 내부의 파라미터, 그리고 L이다.

Test는 Viterbi decoding으로 진행하였다. 디코딩 때 나올 output y 시퀀스는 가장 큰 스코어 S값을 갖는다는 것을 고려해 $y^* = \underset{\tilde{y}\in Y_X}{\operatorname{argmax}}\, s(X,\tilde{y})$를 구하였다. 특히, $\log(\sum_{\tilde{y}\in Y_X} e^{s(X,\tilde{y})})$를 구할 때는 LogSumExp 함수에서 트릭을 이용해 다음과 같은 식(11)을 사용하였고, $x^* = \max\{x_1, x_2, \cdots, x_n\}$으로 표현하였다.

$$\mathrm{LSE}(x_1, x_2, \cdots, x_n) = x^* + \log(\exp(x_1 - x^*) + \cdots + \exp(x_n - x^*)) \quad (11)$$

다음 <Figure 10>과 같이 문장의 단어를 input, Tag를 output으로 놓고 지도학습을 진행하였다.



〈Figure 10〉 A Bidirectional LSTM-CRF model with Lucene Tag Feature 예시

4.3 결과

개체명 인식 방법론의 성능을 평가하는 기준으로 Accuracy와 F1 score를 사용하였다. Accuracy는 전체 결과 중 실제 정답과 같은 판단이 나온 비율을 의미하고, Precision이 실험 결과가 True라고 한 것 중에 실제 True의 비율이며 Recall이 실제 True인 것 중에서 실험 결과도 True라고 한 것의 비율일 때 F1 score는 Precision과 Recall을 기반으로 표현한 것이다. 수식은 다음과 같이 식(12~15)로 각각 표현된다. 그리고 수식에 앞서 사용되는 용어는 <Table 2>를 참고하면 된다.

〈Table 2〉 TP, FP, TN, FN 용어 정리

		실제 정답	
		TRUE	FALSE
실험 결과	TRUE	True Positive(TP)	False Positive(FP)
	FALSE	False Negative(FN)	True Negative(TN)

$$\text{Accuracy(정확도)} = (TP+TN)/(TP+TN+FP+FN) \quad (12)$$

$$\text{Precision(정확률)} = TP/(TP+FP) \quad (13)$$

$$\text{Recall(민감도)} = TP/(TP+FN) \quad (14)$$

$$\text{F1 score} = 2*(\text{Precision}*\text{Recall})/(\text{Precision}+\text{Recall}) \quad (15)$$

Base는 기존 온톨로지에서 인덱스 기반으로 개체명 인식을 하던 방법으로 동음이의어를 처리하지 못하는 것으로 판단하고 성능을 평가하였다. Base method의 경우는 index에 있는 모든 Tag를 뽑아 주기 때문에 F1 score를 구하지 못한다.

4.3.1 동음이의어를 포함한 질의어에 대한 성능

Test set문장 중 동음이의어 포함 문장 <Table 3>에 대해 실험을 진행하였다. 예를 들어 “소녀시대가 부른 곡 장르는?”이라는 질의에서 ‘소녀시대’는 ‘B-MusicArtist’로 인식이 되고, “소녀시대 작곡가가 누구야?”라는 질의에서 ‘소녀시대’는 ‘B-Track’으로 인식이 되는 것을 확인할 수 있는 Data set으로 구성되어 있다.

〈Table 3〉 동음이의어 포함 Data set

D1	동음이의어 포함 문장
Sentence	150
Token	557

결과는 다음 <Table 4>와 같다. Base의 Accuracy는 0.737인데 Bidirectional LSTM-CRF 방법론의 Accuracy는 0.944로 문맥을 고려하여 동음이의어 선택을 높은 성능으로 할 수 있었다. 그런데 L-Bidirectional LSTM-CRF 방법론이 Bidirectional LSTM-CRF 방법론에 비해 성능이 떨어진 이유는 Training set에 특정 Tag가 많을수록 학습도 많이 되기 때문에 Tag feature에서 해당 Tag의 값이 커지는데 동음이의어의 경우에는 이 값이 영향을 미치기 때문에 여러 개의 Tag 중 Tag feature의 값이 큰 Tag가 더 영향력을 끼치는 것이었다.

〈Table 4〉 동음이의어를 포함한 질의어에 대한 성능

		D1
Base	Accuracy	0.737
	F1	-
Bidirectional LSTM-CRF	Accuracy	0.944
	F1	0.750
L-Bidirectional LSTM-CRF	Accuracy	0.827
	F1	0.557

4.3.2 청킹이 필요한 질의어에 대한 성능

Test set문장 중 청킹 포함 문장<Table 5>에 대해 실험을 진행하였다. 예를 들어 ‘발라드’가 “팝 발라드 곡은?”에서는 ‘I-Gerne’로 “발라드 곡은?”에서는 ‘B-Gerne’로 구분이 가능한지, ‘블랙 핑크’라는 ‘B-MusicGroup I-MusicGroup’과 ‘블랙’이라는 ‘B-MusicArtist’, 그리고 ‘핑크’라는 B-MusicArtist’의 구분이 가능한지, 그리고 ‘부르다’가 곡 이름에 속하는 경우일 때와 ‘부르다’라는 서술어와의 구분이 가능한지를 확인할 수 있는 Data set으로 구성되어 있다.

〈Table 5〉 청킹 포함 Data set

D2	청킹포함 문장
Sentence	75
Token	333

결과는 다음 <Table 6>와 같다. Base의 Accuracy는 0.966이며 Bidirectional LSTM-CRF 방법론의 Accuracy는 0.852이다. D2에는 4.3.3절에서 설명하는 학습시키지 않은 단어를 포함한 질의어(D3)와 중복되는 질의가 상대적으로 많았기 때문에 Bidirectional LSTM-CRF 방법론의 성능이 L-Bidirectional LSTM-CRF의 성능보다 상대적으로 떨어졌다.

〈Table 6〉 청킹이 필요한 질의어에 대한 성능

		D2
Base	Accuracy	0.966
	F1	-
Bidirectional LSTM-CRF	Accuracy	0.852
	F1	0.695
L-Bidirectional LSTM-CRF	Accuracy	0.930
	F1	0.909

4.3.3 학습시키지 않은 단어를 포함한 질의어에 대한 성능

Test set문장 중 새로운 단어, Tag 포함 문장 <Table 7>에 대해 실험을 진행하였다. 예를 들어 '이미자', '소방차' 등의 training set에 없었던 단어들로 만든 질의어로 구성되어 있다.

〈Table 7〉 새로운 단어, Tag 포함 Data set

D3	새로운 단어, Tag 포함문장
Sentence	150
Token	557

결과는 다음 <Table 8>와 같다. Base의 Accuracy가 0.930이며 Bidirectional LSTM-CRF 방법론의 Accuracy가 0.688으로 많이 떨어졌으며, L-Bidirectional LSTM-CRF 방법론이 보완할 수 있었다. 특히 F1 score의 값이 큰 차이를 보였다.

〈Table 8〉 학습시키지 않은 단어를 포함한 질의어에 대한 성능

		D3
Base	Accuracy	0.930
	F1	-
Bidirectional LSTM-CRF	Accuracy	0.688
	F1	0.179
L-Bidirectional LSTM-CRF	Accuracy	0.930
	F1	0.765

4.3.4 종합 성능 평가 결과

전체 Test set에 대한 성능을 평가한 결과 <Table 9>로 Base의 정확도는 0.898이 나왔고, Bidirectional LSTM-CRF모델의 정확도가 0.888이 나왔으며 Lucene Tag Feature를 추가한 Bidirectional LSTM-CRF모델의 정확도가 0.909로 더 나은 성능을 보였다.

〈Table 9〉 전체 Test set에 대한 결과

		Test set
Base	Accuracy	0.898
	F1	-
Bidirectional LSTM-CRF	Accuracy	0.888
	F1	0.624
L-Bidirectional LSTM-CRF	Accuracy	0.909
	F1	0.811

4.3.5 종합 속도 측정 결과

L-Bidirectional LSTM-CRF 방법론의 경우 Lucene index DB를 거쳐야 하기 때문에 시간이 오래 걸릴 수 있어 속도 측정 실험을 진행하였다. 결과는 다음 <Table 10>과 같다. 그 결과 Training set 910개의 문장으로 epoch 200을 기준으로 실험 했을 때 Lucene index DB를 거치지 않은 경우와 비교하여 약 2분 4초의 차이를 보였다. Test set 370개의 문장에 대해서는 약 0.1초의 차이를 보였다. Base와 비교를 했을 때는 신경망 모델들이 Train 소요 시간은 오래 걸리지만 Test 소요 시간이 짧은 것으로 나타났다.

〈Table 10〉 전체 Data set에 대한 방법론들의 속도 측정 결과

		소요 시간
Base	Load	3s
	Test	20s
Bidirectional LSTM-CRF	Train	42m 10s
	Test	5.3s
L-Bidirectional LSTM-CRF	Train	44m 16s
	Test	5.4s

5. 결론

기존 온톨로지에서 인덱스 기반으로 개체명 인식을 하던 방법으로는 문맥을 고려하지 못했던 문제가 있었다. 문맥을 고려한 태깅 연구 중 CRF와 Bidirectional LSTM을 결합한 방법이 데이터 특성상 단어 임베딩(Word Embedding)으로 성능향상에 한계가 있고 장기기억 문제를 해결할 수 있지만 최근 입력에 편향될 수 있는 LSTM 기반 모델을 보완하는데 적합하다고 판단하였다. 따라서 실험을 통해서는 동음이의어와 청킹의 경우 문맥을 반영한 Tag 인식이 가능한 것을 확인하였다. 또한, 사용자가 정의한 함수를 변수로 사용할 수 있는 CRF의 특성을 사용하여 데이터의 형태에 맞는 feature를 적용할 수 있었다. 본 연구에서는 새로운 feature로 온톨로지 지식 기반의 특성치를 사용하였다. 그 결과 L-Bidirectional LSTM-CRF 모형을 재학습 시키지 않아도 학습에 포함되지 않은 단어를 포함한 질의에 대한 개체명 인식이 가능함을 확인하였고, 전체적으로 개체명 인식의 성능이 향상됨을 확인할 수 있었다.

따라서 본 연구에서 제안한 L-Bidirectional LSTM-CRF 방법론을 다양한 분야의 온톨로지 지식베이스에 적용하여 개체명 인식 문제를 해결하는데 활용할 수 있을 것으로 보인다.

6. 한계 및 연구 방향

본 연구에서 제안한 L-Bidirectional LSTM-CRF 방법론이 기존 연구 방법에 비해 전반적으로 좋은 성능을 보이나 기존의 Bidirectional LSTM-CRF에 비해 동음이의어를 처리하는 성능이 부족했다. 이는 본 연구의 가설이 학습시킨 질의들이 가지는 개체명의 패턴을 기준으로 모르는 단어의 패턴을 추론할 수 있는 feature를 삽입하는 성질 때문이다. Training Data에 많이 포함되어 있을수록 그 가중치가 선형적으로 증가하는 성질을 가지고 있기 때문에 학습시키지 않은 단어의 개체명을 잘 찾을 수 있지만 반대로 더 많이 나온 개체명에 대한 가중치가 다른 개체명을 가지는 동음이의어의 선택을 막는 것을 알 수 있다.

향후 연구에서는 전반적인 성능을 유지하면서 동음이의어도 잘 찾아낼 수 있도록 가중치를 적용하는 방식을 다양화하여 실험해 볼 필요가 있다.

참고문헌(References)

HUANG, Zhiheng; XU, Wei; YU, Kai. "*Bidirectional LSTM-CRF Models for Sequence Tagging.*" Named entity recognition with bidirectional LSTM-CNNs. arXiv preprint arXiv:1511.08308, 2015.

LAMPLE, Guillaume, et al. "*Neural Architectures for Named Entity Recognition.*" Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360, 2016.

ZHANG, Xiang; LECUN, Yann. "*Text understanding from scratch.*" arXiv preprint arXiv:1502.01710, 2015.

MIKOLOV, Tomas, et al. "*Efficient estimation of word representations in vector space.*" arXiv preprint arXiv:1301.3781, 2013.

KIM, Yoon, et al. "*Character-aware neural language models.*" arXiv preprint arXiv:1508.06615, 2015.

ZHOU, Jie; XU, Wei. "*End-to-end learning of semantic role labeling using recurrent neural networks.*" In: ACL (1). 2015. p. 1127-1137.

GRAVES, Alex; MOHAMED, Abdel-rahman; HINTON, Geoffrey. "*Speech recognition with deep recurrent neural networks.*" In: Acoustics, speech and signal processing (icassp), 2013 ieee international conference on. IEEE, 2013. p. 6645-6649.

GRAVES, Alex; SCHMIDHUBER, Jürgen. "*Framewise phoneme classification with bidirectional LSTM and other neural network architectures*." Neural Networks, 2005, 18.5: 602-610.

LING, Wang, et al. "*Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation.*" In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015. p. 1520-1530.

LAFFERTY, John; MCCALLUM, Andrew; PEREIRA, Fernando CN. "*Conditional random fields: Probabilistic models for segmenting and labeling sequence data*". Proceedings of ICML. 2001.

HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. "*Long short-term memory*." Neural computation, 1997, 9.8: 1735-1780.

ELMAN, Jeffrey L. "*Finding structure in time.*" Cognitive science, 1990, 14.2: 179-211.

Abstract

Improving Bidirectional LSTM-CRF model Of Sequence Tagging by using Ontology knowledge based feature

Seunghee Jin* · Heewon Jang* · Wooju Kim**

This paper proposes a methodology applying sequence tagging methodology to improve the performance of NER(Named Entity Recognition) used in QA system. In order to retrieve the correct answers stored in the database, it is necessary to switch the user's query into a language of the database such as SQL(Structured Query Language). Then, the computer can recognize the language of the user. This is the process of identifying the class or data name contained in the database. The method of retrieving the words contained in the query in the existing database and recognizing the object based on the index does not identify the homophone and the word phrases because it does not consider the context of the user's query. If there are multiple search results, all of them are returned as a result, so there can be many interpretations on the query and the time complexity for the calculation becomes large. In addition, in the existing QA system, chunking has a higher probability of selecting the longest phrase among the sentence component phrases retrieved after obtaining a subset of all cases that can come from the user query. This chunking process has a problem that the algorithm is complicated, but it is not always correct according to the query. To overcome these, this study aims to solve this problem by reflecting the contextual meaning of the query using neural network-based methodology and to identify the problem of the neural network based methodology. Context-sensitive tagging combined with CRF and Bidirectional LSTM can be used to supplement the LSTM-based model, which can limit long-term memory problems, but can be biased toward recent input because of the data nature of word embedding. Therefore, we have solved the disadvantages of the neural network model by introducing the latest technology Bidirectional LSTM-CRF model in Sequence Tagging field. We used reasoning that reflects context by using ontology-based characteristic values for untrained words. In case that untrained words come in, we store the object name recognition tag information obtained from the ontology knowledge base in the Lucene index DB as the

* Information and Industrial Engineering, Yonsei University
** Corresponding Author: Wooju Kim
Information and Industrial Engineering, Yonsei University
50 yonsei-ro, seodaemun-gu, Seoul 03722, Korea
E-mail: wkim@yonsei.ac.kr

ontology knowledge based feature. and In this paper, we propose a neural network model based on ontology knowledge based feature.

Experiments were conducted on the ontology knowledge base of music domain and the performance was evaluated. In order to accurately evaluate the performance of the L-Bidirectional LSTM-CRF proposed in this study, we experimented with converting the words included in the learned query into untrained words in order to test whether the words were included in the database but correctly identified the untrained words. Through experimentation, it was confirmed that tag recognition based on the context of the homophone and chunking is possible. In addition, we could apply the features that match the data type by using the CRF property which can use user defined function as a variable. As a result, it was possible to recognize objects considering the context and can recognize the untrained words without re-training the L-Bidirectional LSTM-CRF model, and it is confirmed that the performance of the object recognition as a whole is improved. Therefore, the proposed L-Bidirectional LSTM-CRF methodology can be applied to the ontology knowledge base of various fields to solve the object name recognition problem. However, the proposed L-Bidirectional LSTM-CRF method performed better than the conventional LSTM-CRF method, but lacked the ability to process homophones compared to the conventional Bidirectional LSTM-CRF. This is due to the fact that the hypothesis of this study inserts a feature that can infer a pattern of unknown words based on the entity name pattern of trained queries. This is because the weight increases linearly with the inclusion in the training data. In future research, it is necessary to experiment with various methods of applying weights so that homophones can be found well while maintaining overall performance.

Key Words : Sequence Tagging, CRF(Conditional Random Field), LSTM(Long Short Term Memory), QA System, Ontology

Received : November 13, 2017 Revised : March 20, 2018 Accepted : March 23, 2018
Publication Type : Regular Paper Corresponding Author : Wooju Kim

저 자 소 개

진 승 희
연세대학교 정보산업공학과 석사
관심연구분야 : 지식그래프, 텍스트마이닝, AI

장 희 원
연세대학교 정보산업공학과 석사
연세대학교 정보산업공학과 박사과정
관심연구분야 : AI, 자연어처리, 지식그래프

김 우 주
연세대학교 정보산업공학과 교수
관심연구분야 : 차세대 웹 기반 기술 및 응용, 웹 서비스 기반 기술 및 응용, 전자상거래(EC) 및 E-Business, 경영정보시스템 및 전문가시스템