seq2seq 주의집중 모델을 이용한 형태소 분석 및 품사 태깅

정의석○, 박전규
한국전자통신연구원, 음성처리연구실
eschung@etri.re.kr, jgp@etri.re.kr

Word Segmentation and POS tagging using Seq2seq Attention Model

Euisok Chung○, Jeon-Gue Park
ETRI, Spoken Language Processing Research Section

요 약

본 논문은 형태소 분석 및 품사 태깅을 위해 seq2seq 주의집중 모델을 이용하는 접근 방법에 대하여 기술한다. seq2seq 모델은 인코더와 디코더로 분할되어 있고, 일반적으로 RNN(recurrent neural network)를 기반으로 한다. 형태소 분석 및 품사 태깅을 위해 seq2seq 모델의 학습 단계에서 음절 시퀀스는 인코더의 입력으로, 각 음절에 해당하는 품사 태깅 시퀀스는 디코더의 출력으로 사용된다. 여기서 음절 시퀀스와 품사 태깅 시퀀스의 대응관계는 주의집중(attention) 모델을 통해 접근하게 된다. 본 연구는 사전 정보나 자질 정보와 같은 추가적 리소스를 배제한 end-to-end 접근 방법의 실험 결과를 제시한다. 또한, 디코딩 단계에서 빔(beam) 서치와 같은 추가적 프로세스를 배제하는 접근 방법을 취한다.

주제어: seq2seq, 주의집중 모델, 형태소 분석, 품사 태깅

1. 서론

한국어 연속어 음성인식이나 기계번역에서 사용되는 언어모델은 형태소 단위의 단어 분할을 통해 기본 어휘를 결정한다. 이는 언어모델의 N-gram 희소성 문제에 기인한다. 또한 단어 분할된 어휘들의 시퀀스를 어절 단위로 복원 시킬 때 품사 태깅 기술을 이용하여 접근할 수 있다. 최근 관련 연구 동향으로는 음절 단위와 품사 정보를 결합하고 기계학습을 통해 형태소 분석과 품사 태깅을 통합하여 진행하는 접근 방법들이 주로 보고 되고 있다[1][2][3]. 본 연구는 음절 기반 형태소 분석 및 품사 태깅을 위해 seq2seq 주의 집중 모델을 이용한다. seq2seq 모델은 인코더과 디코더로 구성되어 있고, 일반적으로 RNN(recurrent neural network)를 기반으로 한다. 학습 단계에서 음절 시퀀스를 인코더의 입력으로 하고, 각 음절에 해당하는 품사 태깅 시퀀스는 디코더의 출력으로 진행한다. 여기서 음절 시퀀스와 품사 태깅 시퀀스는 서로 독립되어 있는데, 대응관계의 모델링을 위해 주의집중(attention) 모델을 이용한다. 본 연구는 사전(dictionary) 정보나 자질 정보와 같은 추가적 리소스를 배제한 end-to-end 접근 방법만을 고려한다. 또한 본 연구는 일반적으로 사용되고 있는 디코딩 단계에서의 빔(beam) 서치를 배제한 접근 방법을 검토해 본다.

2. 관련 연구

음절 기반 한국어 품사 태깅은 미등록어 문제에 대한 처리와 기계학습 적용시 학습될 패러미터의 개수의 감소 시킬 수 있는 접근 방법으로 본 연구를 비롯한 최근 연구의 기본적인 골격은 [1]과 큰 차이가 없다. [1]은 CRF를 이용하여 접근했고, 자질 구조나 디코딩 단계의 자세한 기술은 없다. 반면 Structural SVM을 이용하고 띄어쓰기 문제를 고려한 [2]는 띄어쓰기와 품사 태깅 결합모델의 이용에 대하여 자세히 기술하고 있으며 음절 기반 품사 태깅 성능 정확도 98%의 좋은 결과를 보이고 있다. SVM의 특성상 복잡한 자질 구조를 갖고 있고, 학습 데이터 이외의 명사 사전을 이용한 것으로 기술 되고 있다. 본 연구는 내부 학습 데이터만을 이용한 접근 방법을 제시하는데, 이는 seq2seq모델에 사전 정보를 결합하는 접근 방법이 쉽지 않다는 데도 기인한다. 해당 문제는 향후 연구과제로 남겨둔다. [3]의 경우 [2]의 결과물을 이용하였고, 기분석 사전과 어절 사전을 이용하여 성능을 개선하였다는 보고를 하고 있다.

최근 seq2seq 주의집중 모델을 한국어 형태소 분석 및 태깅에 적용한 연구 사례는 [4]에서 보고 되었다. 본 연구와의 기본적 골격은 차이가 없다. 그러나 학습 단계에서 디코더의 출력으로 어휘정보와 형태소 태그 결합열이 사용된 점과 디코딩 단계에서 빔(beam) 서치를 이용하여 추가적 디코딩 프로세스를 거친다는 점이 차이점이라 할 수 있다. 일반적으로 seq2seq모델은 인코더의 입력 시퀀스를 디코더에서 완벽하게 재현할 수 없다. 해당 문제를 디코딩 빔서치로 해결한 듯 싶다.

3. seq2seq 주의집중 모델

본 연구는 신경망에 기반한 seq2seq 학습 알고리즘을 이용한다. 이에 대한 기초연구는 기계번역에 처음 적용되었다[5]. 여기서 입력 문장과 번역 문장의 쌍에 대하여 입력 문장에 대한 LSTM (Long Short-Term Memory) 인코더와 번역 문장에 대한 LSTM 디코더를 학습하는 접근 방법을 제시하였다. 이는 새로운 입력 문장을 인코더를 통해 문장 임베딩을 하고 해당 임베딩 값을 디코더의 입력으로 하여 번역 문장을 생성하는 end-to-end 접근 방법이다. 기계 번역의 특성상 디코더에서 생성되는 문장

의 구성 어휘는 이와 밀접하게 관련된 인코더의 특정 어휘가 존재한다. 이러한 특징을 모델링하기 위해 [6]는 기계번역에 주의집중 모델을 도입하였다.

[6]에 기술된 수식들을 이용하여 seq2seq 주의 집중 모델을 설명하면 다음과 같다. 입력 시퀀스 $X=(x_1,...,x_n)$에 대한 t시간의 RNN 히든 값은 $h_t=f(x_t,h_{t-1})$로 기술할 수 있다. 입력 시퀀스 전체를 인코딩한 결과는 $c=q(\{h_1,...,h_n\})$가 된다. 여기서 c는 디코더의 입력으로 사용된다. 디코더의 출력 시퀀스를 $Y=(y_1,...,y_m)$라고 했을 때, 어떤 t시간의 y_t의 확률값은 $p(y_t|\{y_1,...,y_{t-1}\},c)$로 기술할 수 있고, 이는 RNN 디코더 $g(y_{t-1},s_t,c)$로 표현하는데 여기서 s_t는 디코더의 히든값을 의미한다. 주의집중 모델은 일반적 인코더 디코더 형태의 seq2seq모델에서 s_t에 인코더의 입력 시퀀스의 각 항목에 대응하는 히든 값들의 가중치 합 (2)를 통합하여 (1)과 같이 모델링한다. 디코더의 히든 값과 인코더의 히든 값의 정렬모델은 (4)로 독립적인 신경망 모델로 구현가능 하다. 학습 단계에서 seq2seq와 통합되어 학습된다.

$$s_t = f(s_{t-1}, y_{t-1}, c_t) \quad (1)$$

$$c_t = \sum_{i=1}^{n} \alpha_{ti} h_i \quad (2)$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{n} \exp(e_{tk})} \quad (3)$$

$$e_{ti} = a(s_{t-1}, h_i) \quad (4)$$

4. seq2seq 입출력 유형

단어 분할 및 품사 태깅을 위한 인코더와 디코더의 입력, 출력 값은 다음과 같이 몇 가지 유형으로 표현할 수 있다.

- 유형1
  X = {그,래,서,#,일,찍,#,잡,니,다}
  Y = {B_maj,I_maj,I_maj,@B_mag,I_mag,@B_pv,B_ef,I_ef}
- 유형2
  X = {그,래,서,#,일,찍,#,잡,니,다}
  Y = {B_maj,I_maj,I_maj,@,B_mag,I_mag,@,B_pv,B_ef,I_ef}
- 유형3
  X = {그,래,서,#,일,찍,#,잡,니,다}
  Y = {B_maj,I_maj,I_maj,B_mag,I_mag,B_pv,B_ef,I_ef}
- 유형4
  X = {그,래,서,일,찍,잡,니,다}
  Y = {B_maj,I_maj,I_maj,B_mag,I_mag,B_pv,B_ef,I_ef}

X는 음절 단위로 구성되고, 공백값은 #으로 처리한다. Y는 형태소 시작 음절 품사는 B를 갖고 이후의 품사는 I로 기술된다. 공백 이후의 형태소 시작 음절 품사는 @B로 시작된다. X와 Y는 각각 동일한 길이는 아니며 인코더와 디코더의 입출력 값으로 학습단계에서 활용된다. 여기서 X와 Y의 구성 요소 간 대응 연결은 주의집중 모델을 통해 학습된다. 실제 테스트 단계에서는 X값만을 입력으로 하고, Y를 생성하게 된다. 유형1은 X에서 음절, 스페이스 정보의 시퀀스로 구성된다. Y는 스페이스 정보가 음절 형태소 태그와 결합된다. 유형2는 Y의 스페이스 정보를 독립적으로 할당한다. 유형3은 Y에서 스페이스 정보를 삭제한다. 유형4의 경우는 입력, 출력 시퀀스 모두에서 스페이스 정보를 제거한다. 이 경우 띄어쓰기 성능 평가 요소가 포함된다고 볼 수 있다.

5. 실험

실험 환경은 텐서플로우의 seq2seq 라이브러리[1]를 이용하여 구현하였다. RNN은 LSTM으로 구현하였다. 학습 형상은 텐서플로우 제공 예제와 동일하게 하였고, LSTM 레이어의 개수와 모델 사이즈, 시퀀스 사이즈를 변경하여 테스트해 보았다.

5.1 seq2seq 입출력 유형 실험

입출력 유형 테스트는 소규모 평가셋으로 진행하였다. 구어체 트위터 형태소 분석 코퍼스를 이용하였는데, 12만 어절로 구성되어 있다. 2만 6천 문장의 학습 문장과 2천 9백 문장의 테스트 문장으로 구성되었고, 개발 셋은 학습 문장의 10%로 선택하였다. 표1은 유형별 성능 결과를 기술한다.

표 1 유형별 실험 결과[2]

	음절COR	음절ACC	문장
유형1	92.50%	91.72%	56%
유형2	92.26%	91.25%	55.3%
유형3	92.50%	91.65%	55.4%
유형4	90.17%	88.14%	46.7%

셀사이즈 64, 임베딩 64, 2-레이어 LSTM을 이용한 seq2seq주의 집중 모델을 이용한 결과에서 유형1이 비교적 좋은 결과를 보였다. 여기서 유형2는 디코더에서 생

1) https://www.tensorflow.org/versions/r0.10/tutorials/seq2seq/index.html

2) N=평가레이블수, I=삽입오류, D=삭제오류, S=변경오류,

$$Correct = \frac{N-D-S}{N} \times 100\%, \; Accuracy = \frac{N-D-S-I}{N} \times 100\%$$

성된 시퀀스에서 스페이스를 배제한 결과이다. 유형4의 경우는 띄어쓰기 정보가 배제된 상태라 스페이스 정보의 중요성을 확인함과 동시에 띄어쓰기 오류 대처가 가능하다는 점을 확인할 수 있었다.

5.2 세종 코퍼스 및 구어체 실험

첫 번째 평가 셋은 세종 형태소 분석 코퍼스 110만 어절을 워드 분할 형식으로 변환하였고, 총 10만 4천 문장의 학습데이터, 1만1천 문장의 개발데이터, 1만2천 문장의 평가 데이터로 랜덤하게 선택하여 구성하였다. 두 번째 평가셋은 트위터, 이메일, 인터넷 게시판 등에서 추출한 구어체 문장으로 구성된 91만 어절을 워드 분할 형식의 형태소 분석 코퍼스로 변환하였고, 문장 개수는 세종 평가 셋과 유사하게 하였다. 기호를 제외한 영어 대소문자, 한자 등은 대표 문자로 변환하였다. 각 구성문장의 길이는 공백을 포함한 음절 시퀀스 50개로 제한하였고, 강제적으로 분할하여 문장을 구성하였다.

표 2 실험 결과

	음절COR	음절ACC	문장
본 연구			
세종(유형1)	94.32%	93.96%	43.5%
세종(유형4)	92.47%	91.28%	32.7%
구어체(유형1)	95.77%	95.56%	60.7%
기존 연구			
[2]세종(유형1)	98.02%	-	-
[2]세종(유형4)	96.99%	-	-
[4]세종(유형1)	-	-	60.62%

표2는 실험 결과로 음절단위, 문장단위 기준의 평가를 진행하였다. seq2seq 주의 집중 모델은 3-레이어, 64사이즈의 LSTM으로 인코더, 디코더를 구성하였다. 실험 결과는 기존 연구 결과들 [2][3]의 성능인 세종 평가셋 음절 단위 성능인 98.02% 수준에 못 미치는 결과인 94.32%를 보였고, 유사한 구조를 지닌 [4]의 문장 단위 평가 60.62%에 비해 부족한 결과인 43.5%를 보였다. 평가셋 자체가 동일하지 않지만 [4]와의 차이점인 디코더 출력 시퀀스에 음절 시퀀스 정보를 포함하고, 이를 이용한 디코더 빔서치 프로세스의 타당성을 확인할 수 있었다.

구어체 수준의 영역에는 95.77%의 음절 정확도와 문장 정확도 60.7% 수준의 결과를 보였다. 또한 사전 정보와 같은 외부자질을 활용하지 않은 결과라 개선의 여지가 충분한 접근 방법이라 보고, 현 상태에서 후처리를 보완한다면 어느 정도 활용 가능성을 확인할 수 있었다.

6. 결론

본 논문은 사전정보, 자질정보, 디코딩 프로세스를 배제한 상태에서 end-to-end 접근 방법인 seq2seq 주의집중 모델이 형태소 분석 및 태깅에 활용 가능한지 검토해 보았다. 구어체 영역의 경우, 입출력 학습 데이터 작업만으로 쉽게 실 영역에서 활용 가능한 형태소 분석 및 태거를 확보할 수 있다고 판단 되었다. 향후 추가적 작업으로 seq2seq 모델의 인코더 단계나 디코더 단계에서 사전정보와 자질정보 등의 활용 가능한 리소스를 모델에 통합하는 접근 방법 연구 등이 가능하리라 본다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였습니다. [R0126-15-1117, 언어학습을 위한 자유발화형 음성대화처리 원천기술 개발]

참고문헌

[1] 심광섭, "형태소 분석기 사용을 배제한 음절 단위의 한국어 품사 태깅", 인지과학, 제22권, 제3호, pp.327-345, 2011.

[2] 이창기, "Structural SVM을 이용한 한국어 띄어쓰기 및 품사 태깅 결합 모델", 정보과학회논문지:소프트웨어 및 응용, 제40권, 제12호, pp.826-832, 2013.

[3] 이충희, 임준호, 임수종, 김현기, "기분석사전과 기계학습 방법을 결합한 음절 단위 한국어 품사 태깅", 정보과학회논문지, 제43권, 제3호, pp.362-369, 2016.

[4] 이건일, 이의현, 이종혁, "Sequence-to-sequence 모델을 이용한 한국어 형태소 분석 및 품사 태깅, 한국컴퓨터종합학술대회(KCC), pp.693-695, 2016.

[5] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).