seq2seq을 이용한 형태소 분석

양봉석

Song, Park(2019)_Korean morphological analysis with tied sequence to sequence multi-task model

1. Introduction

- 한국어 형태소 분석기의 목적: 어절을 형태소 단위로 분석하는 것 + 분석된 단위에 POS tag를 부착하는 것.
- 전통적인 방법: 형태소 분석 후 POS tagging하기(Lee and Rim, 2009; Na, 2015; Choi et al. 2016; Matteson et al., 2018; Song and Park, 2018)
    - 단점:
        1. 형태소 처리 과정에서의 오류가 POS tagging에도 영향을 미친다.
        2. 형태소 처리와 POS tagging 간의 mutual information을 모델링 하기가 어렵다.
- 한국어에서 형태소 처리와 POS tagging은 서로 영향을 미친다. 즉, 형태소 분석을 정확하게 하면 POS tagging도 정확해진다. 반대로 정확한 POS tagging이 형태소 분석을 더 정확하게 만들어주기도 한다.

Figure 1: Korean morphological analysis of a sentence "나는 하늘을 나는 새를 봤다" of which meaning is "*I saw a bird flying in the sky*". Correct morpheme analysis helps predicting POS tags (blue dotted arrows) while POS tagging affects morpheme analysis (red arrows). Orange morphemes in the morpheme sequence are recovered morphemes. Best viewed in color.

    - ‘나는(I)’와 ‘나는(flying)’의 차이를 잘 구별하기 위해서는 **형태소 분석과 POS tagging이 동시에 학습**되어야 한다.
- 따라서 이 논문은 형태소 처리와 POS tagging을 동시에 학습하는 모델을 만들 것이다.
- 이 모델은 형태소 처리와 POS tagging을 개별 task로 간주함.
- 각 task는 seq2seq multi-task framework로 jointly train됨.
- 이 모델의 주요 특징: 형태소 시퀀스를 만드는 것과 POS tag를 부착하는 것 모두 단일한 decoder를 갖는다. 이로 인해 디코더는 이 두 task의 general representations를 공유하며, 형태소와 POS tag 간의 1:1 매핑을 하게 된다.
    - 형태소 분석은 pointer generator network(See et al., 2017) 이용
        - pointer generator network: 입력 텍스트에서 토큰을 가져올지 생성할지를 결정하고 토큰의 반복을 피하고자 이전 토큰들의 분포를 고려하여 토큰을 생성하는 모델
    - POS tagging은 CRF network 사용

2. Korean Morphological analysis

2.1. Morpheme processing and POS tagging

- 한국어 형태소 분석의 특징: input token의 길이와 output token의 길이가 다르다.
    - `축구선수` → `축구`, `선수`

2.2. Linguistic Unit in Morphological Analysis

- 어절: 한국어 문장의 띄어쓰기 단위.
- seq2seq 모델에 ‘어절’ 단위를 input으로 넣는 것은 부적절함. (계산복잡도 증가)
- 어절은 음절로 이루어지므로 input 단위를 음절로 하면 계산복잡도를 낮출 수 있음. 따라서 이 논문에서는 음절 단위를 input 단위로 설정함.

3. **Tied Sequence to Sequence Multi-task model**

- The proposed model is based on the sequence-to sequence model with attention (Bahdanau et al., 2015) and extends the model for multi-task learning similarly to the work of Anastasopoulos and Chiang (2018)
- 모델의 구성

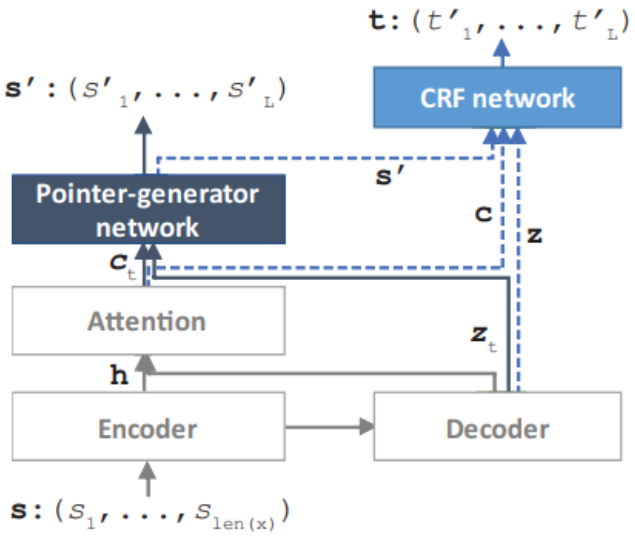

Figure 2: The proposed model for Korean morphological analysis based on a Tied sequence-to-sequence multi-task model. Since syllable is adopted as a unit for the proposed model, the input for the encoder is a syllable sequence s. For clarity's sake, there are some dependencies not shown.

- recurrent encoder
  - bidirectional LSTM
  - input sequence **x**를 hidden state의 sequence로 인코딩한다.
  - s: 음절(syllable)
- attention
  - encoder hidden states **h**를 attention weights를 통해 context vector **c**로 변환함. 이때 context vector는 형태소 처리, POS tagging의 정보를 모두 담고 있음.
- recurrent decoder
  - unidirectional LSTM
  - 형태소 처리 task, POS tagging task라는 두 task를 하나의 디코더로 처리함.
  - 두 task가 하나의 동일한 decoder states를 공유하기 때문에, 별도의 추가적인 작업 없이도 형태소와 POS tag 간 일대일 매핑이 보장된다.
  - 단방향 LSTM이 decoder hidden state **z**를 계산한다. 그리고 매 시점 **t**마다 **z_t**가 이전 음절 **s'_t1**의 임베딩, 이전 시점의 decoder state **z_t-1**, 현

재 시점의 context vector **c_t** 으로부터 계산된다.

- 이후 task-dependent network가 동일한 **z**를 가지고 각각의 task를 푼다.

- task-dependent network
  - 그림의 굵은 화살표: **morpheme processing**
    - 음절 **s'_t**가 매 시점 **t**마다 생성됨.
    - 대부분의 음절들은 input 음절로부터 복사되나, 일부는 형태소의 기본형(base form)을 복원하기 위해 vocabulary로부터 새롭게 생성된다. (예: `하늘을 [나는`] → `날 + 는`
      - 이 작업을 위해 pointer-generator network(See et al., 2017) 도입. pointing을 통해 음절들을 복사하고, fixed vocabulary로부터 음절들을 생성한다.
      - 이 방식의 장점: 한자, 특수문자와 같은 non-Korean 문자를 처리할 수 있음.
  - 그림의 점선: **POS tagging**
    - pointer-generator netword를 통과한 다음 CRF network를 통과함.
    - 형태소 처리로부터 생성된 형태소 시퀀스가 주어지면, CRF network가 tag들 간의 global dependency를 고려해서 POS tag를 예측한다.
    - 더불어 이 연구에서는 CRF network가 디코더의 hidden state에 attention할 수 있도록 skip-connections(He et al., 2016)를 사용하였음.
    - 구체적으로, pointer-generagor network로부터 생성된 음절 시퀀스가 벡터로 변환되고, 이 벡터들이 decoder state **z**, context vector **c**와 concatenate된다.
    - 그러고 나서 concatenate된 벡터가 CRF 층으로 전달되어 POS tag를 예측함.
    - 단, input은 음절 단위인데 반해 POS tag는 형태소 단위이므로, 하나의 POS tag는 여러 개의 음절에 걸칠 수 있다. 그러므로 형태소 경계를 식별하기 위해 형태소의 끝 지점에 special symbol을 생성시킴.
  - 모델은 각 task의 conditional log-likelihood의 weighted sum을 최대화하는 방향으로 훈련됨.

4. Experiments

4.1. Experimental Settings

- 데이터셋: 세종 코퍼스

Information	Training	Development	Test
Sentences	197,508	5,000	50,631
Eojeols	2,674,571	97,292	694,524
Morphemes	5,952,985	203,244	1,542,483
Avg. eojeols	13.54	19.46	13.72
No. of POS tags	42		

Table 2: Simple statistics on the data set used.

- 음절 embedding size, hidden size: 100
- LSTM layers: 3
- batch size: 128
- gradient normalization
- metric:
  - morpheme level: F1-measure
  - eojeol level: accuracy

4.2. Experimental Results

Methods				F1-score	Acc (%)
Model	Morpheme processing	POS tagging	Type		
Single-Task	Pointer-Generator			97.36	95.57
Multi-Task	Generator	Generator	Non-cascade	97.10	95.09
	Generator	CRF	Cascade	97.25	95.30
	Generator	CRF	Tied	97.31	95.49
	Pointer-Generator	Generator	Non-cascade	97.30	95.45
	Pointer-Generator	CRF	Cascade	97.40	95.61
	Pointer-Generator	CRF	Tied	**97.43**	**95.68**
Pipeline	Na (2015) (CRF → CRF → Post-processing)			97.21	95.22
	Song and Park (2018) (Generator → CRF)			97.27	95.29
	khaiii (CNN → Post-processing)			94.88	91.86

Table 1: Performances of Korean morphological analyzers

- Non-cascade: 각 task가 decoder만 공유하고, 두 task 간 직접적인 연결은 없음.

- cascade: POS tagging network가 morpheme processing network와 연결되어 있으나 decoder로부터의 skip connection이 없음.
- Tied: 이 논문의 제안 모델. 직접적인 연결 O, skip connection O
- Tied > cascade > non-cascade
- 기존의 pipeline 모델보다 성능이 더 좋다.

4.3. Error Analysis

Type	Percentage
Morpheme segmentation	5.8%
POS tagging	39.3%
Morpheme recovery	54.9%

Table 3: The ratio of different error types.

- Morpheme segmentation error: 잘못 분절된 비율
  - 주로 합성명사(compound noun)
- POS tagging error: 형태소 분석은 정확하나 POS tag가 잘못 예측된 비율
  - 명사, 고유명사에서 주로 발생
- Morpheme recovery error: 어절로부터 형태소를 잘못 복원한 비율
  - 가장 높았음(54.9%)

5. 결론
   - morpheme processing + POS tagging을 함께 학습한 seq2seq 모델이 기존 형태소 분석 모델보다 성능이 더 좋았다.

Reference

Song, H. J., & Park, S. B. (2019, November). Korean morphological analysis with tied sequence-to-sequence multi-task model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 1436-1441).