양방향 LSTM과 선형체인 CRF를 이용한 복합명사 분해

이현영, 강승식
국민대학교 컴퓨터공학과
e-mail:le32146@gmail.com, sskang@kookmin.ac.kr

Compound Noun Decomposition by using Bi-LSTM and Linear-chain CRF

Hyun-Young Lee, Seung-Shik Kang
Dept of Computer Science, Kookmin University

요 약

복합명사 분해 문제를 태그열 부착 문제로 정의하고 음절 임베딩과 딥러닝을 이용하여 복합명사를 분해하는 방법을 제안한다. 임베딩 방식으로는 음절 단위로 복합명사에 출현한 음절들을 벡터 공간에 표현하고 양방향 LSTM과 선형체인(linear-chain) CRF를 이용하여 복합명사 분해 태그를 부착하여 복합명사를 단위명사들로 분해하였다.

1. 서론

한국어의 복합명사 분해에 관한 연구는 규칙기반 방식과 말뭉치 기반의 확률 및 통계적 방식으로 나누어진다. 규칙기반 방식은 복합명사의 음절 길이에 따른 선호 음절 패턴 규칙을 이용하는 방식으로 단위명사 사전, 접사 사전, 한국어 복합명사의 구조적 특성을 기반으로 복합명사를 분해한다[1,2].

강승식(1998)은 형태소 분석 결과로 추정된 4음절 복합명사의 2+2, 5음절 복합명사의 3+2, 2+3 등의 음절 패턴 유형과 두 가지 예외 규칙을 사용하여 복합명사를 분해하는 방법을 제안하였다. 이러한 규칙 기반의 방식은 음절 길이의 제한이 없이 결합이 가능한 모든 복합명사를 대상으로 하기에는 어려움이 있다.

확률 기반의 방법은 음절 n-gram 빈도를 이용하는 방식으로 복합명사를 구성하는 단위명사의 위치별 명사빈도 데이터를 이용하여 복합명사를 분해한다. 심광섭(1997)은 이웃하는 두 음절간의 합성 상호 정보(composite mutual information)를 이용하여 띄어쓰기가 전혀 되어 있지 않은 복합명사를 단위명사로 분해하는 알고리즘을 제시하였다[3].

복합명사 분해는 분해 기준에 따라 중의성으로 인하여 복합명사의 구문적 불일치 문제가 발생하며, 사전 기반의 복합명사 분해 방식은 사전에 등록되지 않은 미등록어인 단위명사를 인식하기 어려운 점이 있다. 본 논문에서는 음절을 벡터 공간에 표현하는 음절 벡터와 딥러닝 기법으로 음절 단위의 정보를 이용하는 복합명사 분해 방법을 제안하였다.

2. 딥러닝을 이용한 복합명사 분해 방법

복합명사 분해 문제를 BI 태그열(tag sequence) 생성 문제로 정의하고 복합명사의 각 음절에 분리 태그를 부착한 후에 그림 1과 같이 음절 사이에 공백이 없는 복합명사를 단위명사들로 분해한다. 복합명사 분해 태그는 B(begin)와 I(inside)를 사용한다.

복합명사는 두 개 이상의 명사들이 결합되지만 명사를 구성하는 성분은 음절 단위이고, 명사를 구성하는 음절 종류는 한정적이다. 따라서 사전에 미등록 단위명사를 구성하는 음절이 사전에 등록된 단위명사를 구성하는 음절로 사용되는 경우가 단위명사 자체로 사전에 존재하는 확률보다 높다. 본 논문에서는 복합명사들에 대해 음절 unigram과 bigram 형태의 음절 사전을 구축하고 음절 임베딩과 딥러닝 기법을 이용하여 복합명사를 단위명사로 분해한다.

딥러닝 모델은 단어를 표현하기 위해서 단어 또는 음절을 연속적인 벡터 공간에 표현하는 임베딩 기법을 사용하는데 음절 unigram과 bigram을 연속적인 벡터 공간에 표현한다[4]. 양방향 LSTM-CRF는 순차적인 태그열 분류에서 사용되는 모델이다[5]. 양방향 LSTM은 현재 입력 자질(feature)과 과거, 미래의 자질 정보간의 의존성을 연관시켜 새로운 자질 벡터를 생성한다. 선형체인 CRF는 양방향 LSTM과 전방향 신경망(feedforward neural network)을 통해 생성된 새로운 태그와 해당 입력의 출력 태그열을 함께 고려하여 최적 태그열을 예측한다.

복합명사를 구성하는 각 unigram 음절과 bigram 음절의 벡터 표현을 위해 그림 1과 같이 음절 unigram 벡터와

bigram 벡터를 양방향 LSTM를 이용하여 인코딩(encoding)한 새로운 자질 벡터를 음절 unigram과 bigram에 대응되는 순서로 결합하여 새로운 자질 벡터(feature vector)를 생성하였다. 이 새로운 자질 정보를 전방향 신경망의 입력으로 사용하여 각 unigram 음절에 해당하는 BI 태그 클래스의 점수(score)를 계산하고 선형체인 CRF를 통해 최적의 태그열을 예측하여 복합명사를 단위명사로 분해한다.

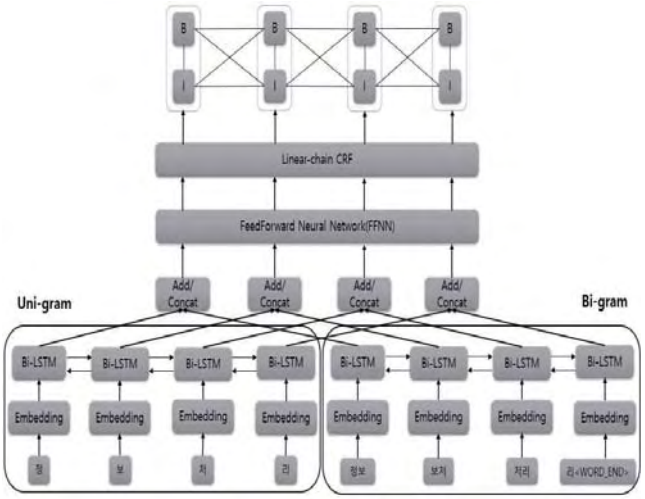

그림 1. Bi-LSTM-CRF with bigram embedding

3. 실험 및 성능 평가

복합명사 분해 실험 및 평가를 위한 말뭉치 데이터는 "차세정 언어처리 경진대회 2018"의 복합명사 분리 태스크(task)에서 제공하는 복합명사 말뭉치를 사용하였다. 말뭉치 크기는 총 2,889,709개의 복합명사로 단위명사 320,532개(중복 단위명사 미포함)로 구성되어 있다. 복합명사 분해 학습 및 평가를 위해 2,889,709개의 복합명사를 2,600,738개의 학습 데이터와 288,971개의 테스트 데이터로 나누어 학습 및 평가를 수행하였다.

표 1은 학습 및 평가를 위한 음절 임베딩, 양방향 LSTM의 출력 연산의 종류, 음절 임베딩 크기 등을 다르게 구성한 모델의 종류이다.

표 1. 모델 종류

모델	음절 임베딩	양방향 LSTM 출력 연산 종류	임베딩 크기
1	Unigram	Add	250
2			300
3	Unigram + Bigram	Add	250
4			300

*Epoch: 5, 10, 15, 20
*Learning Rate: 0.001
* Batch Size: 10, 20

모델의 정량적 성능 평가를 위해 식 (1)~(3)과 같이 복합명사 분해 태그 정확도(accuracy), 어절 재현율(word recall)과 공백 재현율(spacing recall)을 사용하였다. 표 2는 모델 종류별 성능 평가 및 비교 결과이다.

$$Accuracy = \frac{The\ Predicted Correct\ Tags}{The\ Actual\ Whole\ Tags} \times 100 \quad (1)$$

$$Word\ Recall = \frac{The\ Predicted\ Correct\ Words}{The\ Actual\ Entire\ Words} \times 100 \quad (2)$$

$$Spacing\ Recall = \frac{The\ Predicted\ Correct\ Spacing}{The\ Actual\ Entire\ Spacing} \times 100 \quad (3)$$

표 2. 복합명사 분해 성능 평가 (단위: %)

모델	복합명사 태그 정확도	단위명사 재현율	공백 재현율
1	93.57	86.04	87.10
2	93.52	86.23	87.54
3	97.35	94.05	94.76
4	97.31	93.90	94.56

4. 결론

복합명사 분해를 순차적인 태그 부착 문제로 정의하고 복합명사를 단위명사로 분리하기 위해 음절 단위 임베딩 기법으로 양방향 LSTM과 선형체인 CRF를 이용하는 딥러닝 모델을 제안하였다. 실험 결과로 음절 unigram 임베딩 기법보다 음절 unigram과 음절 bigram을 함께 연속적인 벡터 공간에 표현했을 때 성능이 더 높게 나타났다.

감사의 글

본 연구는 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2017M3C4A7068186).

참고문헌

[1] S. Kang, "A decomposition algorithm of Korean compound nouns," Journal of KISS(B): Software and Applications, Vol.25, No.1, pp.172-182, 1998.

[2] Y. Hoon Lee et al. "Korean compound noun decomposition and semantic tagging system using user-word intelligent network," The KIPS Transactions: Part B, Vol.19, pp.63-76, 2012.

[3] K. Shim, "A compound noun segmentation using composite mutual information," Journal of KISS(B): Software and Applications, Vol.24, No.11, pp.1307-1317, 1997.

[4] D. Lee, Y. Lim and T. Kwon, "Morpheme-based efficient Korean word embedding," Journal of KIISE, Vol.45, No.5, pp.444-450, 2018.

[5] Z. Huang, W. Xu, and K. Yu. "Bidirectional LSTM-CRF models for sequence tagging," arXiv preprint arXiv:1508.01991. 2015.