심층신경망 기반 2단계 한국어 자동 띄어쓰기 모델

최기현○, 김시형, 김학수
강원대학교, 컴퓨터정보통신공학과
pluto32@kangwon.ac.kr, sureear@kangwon.ac.kr, nlpdrkim@kangwon.ac.kr

Two Step Automatic Korean Word Spacing Model Based on Deep Neural Network

Gihyeon Choi○, Sihyung Kim, Harksoo Kim
Kangwon National University Computer and Communication Engineering

요 약

자동 띄어쓰기는 띄어쓰기가 되어있지 않은 문장에 대하여 띄어쓰기를 해주거나, 문장에 있는 잘못된 띄어쓰기를 교정하는 것을 말한다. 기존의 자동 띄어쓰기 연구는 주로 모든 음절을 붙인 후 새로 띄어쓰기 태그를 입력하는 방법을 사용하여 사용자가 입력한 올바른 띄어쓰기 정보를 활용하지 못하였다. 따라서 본 논문에서는 모두 붙여 쓴 문장에 공백을 넣어주는 띄어쓰기 삽입 모델과 사용자의 입력 정보를 이용하여 문장의 띄어쓰기 오류를 교정해주는 오류교정 모델이 결합된 통합모델을 제안한다. 제안된 모델은 에러율 10%일 때 F1-score가 98.85%까지 향상되었다.

주제어: 자동 띄어쓰기, BiGRU-CRFs, 사용자의 입력 정보, 오류교정

1. 서론

띄어쓰기는 언어의 문자 표기 시 단어 또는 의미 단위 사이에 공백을 넣어 간격을 벌리는 표기법을 말한다. 옳지 않게 띄어쓰기 된 문장은 글의 가독성을 떨어뜨리고 이를 자연 언어 처리 응용에 사용할 경우 오류를 전파시키는 등 여러 문제를 발생시킨다. 따라서 자연 언어 처리 응용에 있어서 띄어쓰기 전처리 작업은 필수적이다[1]. 자동 띄어쓰기는 띄어쓰기가 생략된 문장을 자동으로 띄어쓰기 해주거나, 일부 잘못된 띄어쓰기가 되어있는 부분을 고쳐주는 것을 말한다. 자동 띄어쓰기에 관한 기존의 연구들은 주로 문장의 공백을 모두 제거한 문장을 입력으로 사용하는 것에 집중되었다. 하지만 입력 문장의 띄어쓰기 정보를 모두 생략한다면 사용자가 올바르게 입력한 띄어쓰기 정보까지 삭제될 수 있다. 따라서 본 논문에서는 띄어쓰기가 생략된 문장을 자동 띄어쓰기 해주는 띄어쓰기 삽입 모델과 그 출력과 사용자의 입력 정보를 이용하여 띄어쓰기 오류를 교정해주는 띄어쓰기 오류교정 모델이 합쳐진 통합모델을 제안한다.

2. 관련 연구

기존의 자동 띄어쓰기에 관한 연구는 규칙 기반 방식과 통계 기반 방식이 있다. 규칙 기반 방식은 여러 언어학적 자원을 요구하며 어휘 지식, 사전정보를 이용한 휴리스틱을 적용하는 방식이다. 규칙 기반 방식은 구현이 까다롭고 미등록어 처리에 취약한 단점이 있다. 통계 기반 방식은 대량의 말뭉치를 이용하여 음절들의 확률 정보를 학습하여 자동 띄어쓰기하는 방식을 말한다. 통계 기반 방식은 대량의 학습데이터 외에 다른 언어학적 자원이 필요하지도 않고 미등록어 처리에서도 규칙 기반 방식에 비해 비교적 높은 성능을 보이고 있다. 이와 같은 이유로 최근 통계 기반 방식의 여러 연구가 진행되었다[2]. 기존의 연구들은 자동 띄어쓰기 문제를 순차적 레이블(sequence labeling)링 문제로 보고 접근하였다. [3]은 순차적 레이블링 문제에 적합한 CRFs(Conditional Random Fields)를 이용한 자동 띄어쓰기 모델을 제안하였다. [4]는 SVM(Support Vector Machine)을 순차적 레이블링 문제에 적용할 수 있도록 확장한 Structural SVM 기반 모델을 제안하였다. 최근에는 자연 언어 처리 분야에 높은 성능을 보이고 있는 심층신경망(Deep Learning) 기법을 사용한 연구가 진행되었다. [5]는 순환신경망(Recurrent Neural Network)의 기울기 소실문제(Vanishing Gradient Problem)를 개선한 GRU(Gated Recurrent Unit)에 CRFs를 결합한 GRU-CRFs를 사용한 모델을 제안하였다. [6]은 Structural SVM을 이용한 띄어쓰기 모델을 기반으로 사용자가 입력한 문장의 띄어쓰기 정보를 최대한 보존하며 자동 띄어쓰기 해주는 모델을 제안하였다.

본 논문에서는 순환신경망을 개선하여 양방향 상태 정보를 모두 이용하는 BiGRU-CRFs(Bidirectional GRU-CRFs)[7]를 기반으로 문장을 자동으로 띄어쓰기 해주고, 사용자의 띄어쓰기 정보를 이용하여 오류를 교정해줌으로써 성능을 향상시키는 모델을 제안한다.

3. 다층 Bidirectional-GRUCRFs를 이용한 한국어 자동 띄어쓰기 모델

본 논문은 자동 띄어쓰기 문제를 순차적 레이블링 문제로 접근하여, 입력 문장의 각 음절에 태그(B, I)를 부착하는 방식을 사용한다. 어절의 시작 음절의 경우 'B', 그 외의 음절에는 'I'를 부착한다. 예를 들어

“나는 한국인입니다.” 라는 문장에 대한 태그는 표 1과 같다.

표 1 “나는 한국인입니다”의 태그 예시

X	나	는	한	국	인	입	니	다	.
Y	B	I	B	I	I	I	I	I	I

3.1 띄어쓰기 삽입모델과 오류교정 모델

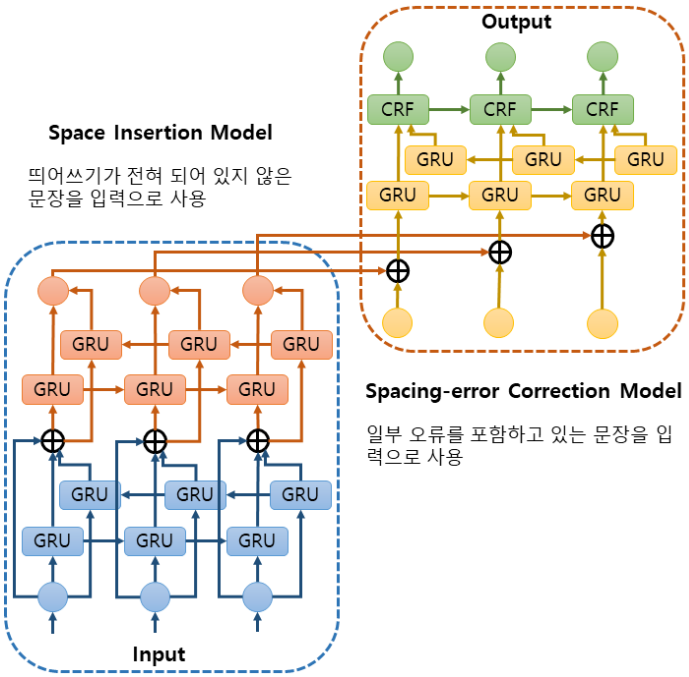

그림 1 모델 전체 구조도

그림 1은 본 논문에서 제안하는 모델의 전체 구조도이다. 본 논문에서 제안하는 모델은 띄어쓰기 삽입 모델과 띄어쓰기 오류교정 모델이 합쳐진 통합모델 구조이다. 삽입 모델과 오류교정 모델 모두 순차적 레이블링 문제에 높은 성능을 보인 BiGRU-CRFs를 기반으로 한다. 띄어쓰기 삽입 모델은 띄어쓰기가 생략된 문장을 입력으로 띄어쓰기가 된 문장을 출력하고 2개의 BiGRU(Bidirectional GRU) 계층으로 구성되어 있다. 입력은 음절 기준이며, 해당 음절에 대한 임베딩 벡터(Embedding vector)와 각 음절에 대한 자질을 사용한다. 삽입 모델의 두 번째 계층은 각 음절의 역할을 명확히 구분하기 위해 첫 번째 계층에서 출력된 은닉 상태(Hidden state)를 입력 음절의 임베딩 벡터와 연결하여 사용한다. 띄어쓰기 교정 모델은 삽입 모델과 동일한 입력에 추가로 삽입 모델의 은닉 상태와 사용자의 띄어쓰기 정보를 연결(Concatenate)하여 사용한다. 이 때, 사용자의 띄어쓰기 정보는 실제 정답에 임의로 오류를 발생시킨 결과를 사용한다. 오류 교정 모델의 경우 삽입 모델의 은닉 상태를 입력으로 사용하기 때문에 한 개의 계층으로만 구성되어 있다.

3.2 명사 사전을 이용한 자질 표현

문장에 명사가 있을 때, 일반적으로 해당 명사의 앞에 띄어쓰기가 존재한다. 이러한 이유로 명사 사전을 이용한 자질을 사용한다. 입력으로 사용되는 기준 음절을 x_t라 할 때, x_{t-2}에서 x_{t+2}까지 5음절에 대하여 n-gram(n=2,3,4,5)을 적용한다. n-gram을 적용하여 추출한 글자가 명사 사전에 있을 경우 1, 없을 경우 0으로 표현한다.

표 2 음절 n-gram 추출 예시

“나는 학교에 갑니다.” 에서 ‘학’ 의 경우	
n = 2	나는, 는학, 학교, 교에
n = 3	나는학, 는학교, 학교에
n = 4	나는학교, 는학교에
n = 5	나는학교에

표 2는 “나는 학교에 갑니다” 라는 문장이 있을 때 “학” 음절의 n-gram 추출 예시이다.

4. 실험 및 결과

4.1 실험 환경

본 논문에서는 학습데이터로 뉴스말뭉치(253,138문장)와 세종말뭉치(257,071문장)를 사용하여 학습하였다. 평가데이터로는 ETRI 품사 태그 말뭉치(27,854문장)를 사용하였다.

오류교정 모델의 입력으로 사용되는 사용자의 띄어쓰기 정보는 실제 정답에 임의로 오류를 발생시켜 사용하였다. 본 논문에서는 특정 음절에 대한 태그를 무작위로 “B” 또는 “I” 태그를 선택하여 변경한다. 표 2의 문장을 기준으로 한 예시는 다음과 같다.

표 3 에러율 30% 예시

X	나	는	한	국	인	입	니	다	.
Y	B	I	I	B	I	I	B	I	I

바뀌는 태그는 문장의 첫 음절을 제외한 나머지 음절 중에서 무작위로 선택되며, 변경한 태그 개수에 따라 에러율이 결정된다. 에러율이 30%인 경우 (전체 음절 개수 -1)*0.3 만큼의 태그가 변경되었다는 것을 의미한다. 다만, 에러율 100%인 경우에는 모든 음절을 붙여 쓴 문장을 사용하였다.

4.2 실험 결과

표 4 에러율에 따른 모델 성능

	정확률	재현율	F1-score
에러율 10%	98.66	98.51	98.58
에러율 30%	97.42	97.54	97.48
에러율 50%	96.61	97.06	96.83
에러율 70%	96.08	96.73	96.41
에러율 100%	95.69	96.5	96.1

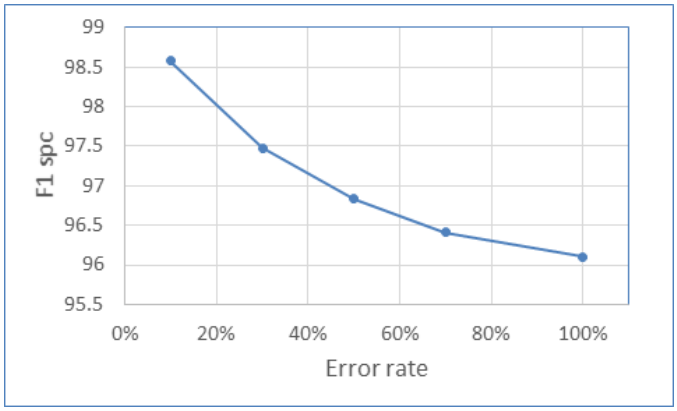

그림 2 에러율에 따른 성능 변화 그래프

표 4는 각각의 에러율에 따른 모델의 성능을 보여준다. 그림 2는 에러율에 따른 성능 변화를 좀 더 잘 볼 수 있도록 그래프로 표현한 것이다. 그림 2를 보면 에러율이 떨어질수록 그에 비례하여 성능이 증가하는 것을 볼 수 있으며, 에러율이 10%인 경우에는 F1 점수가 98.58%까지 향상되었다.

표 5 성능표(복합명사 고려)

	정확률	재현율	F1-score
에러율 10%	98.89	98.82	98.86
에러율 30%	98.02	98.21	98.11
에러율 50%	97.55	97.95	97.75
에러율 70%	97.23	97.75	97.49
에러율 100%	97.06	97.65	97.35

한국어의 경우 복합명사는 붙여 쓰는 경우, 단위 명사별로 띄어 쓰는 경우 모두를 허용한다[8]. 표 5는 복합명사 현상을 고려한 성능이다.

5. 결론

본 논문에서는 띄어쓰기 정보가 생략된 문장뿐만 아니라 사용자의 띄어쓰기 입력 정보를 이용하여 문장의 포함된 일부의 오류를 찾아 교정해주는 띄어쓰기 모델을 제안하였다. 실험 결과 사용자의 입력 문장이 정확할수록 높은 성능을 보였으며 F1 점수가 96.1%에서 98.58%까지 향상되었다. 향후 연구로는 실제 띄어쓰기 오류 데이터를 구축하여 실험하고, 다른 딥러닝 모델을 이용한 자동 띄어쓰기 모델을 연구할 예정이다.

감사의글

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.2016R1A2B4007732)

참고문헌

[1] 김선우, 최성필, “Bidirectional LSTM-CRF 기반의 음절 단위 한국어 품사 태깅 및 띄어쓰기 통합모델 연구”, *정보과학회논문지*, 제45권, 제8호, 792-800, 2018.
[2] 송영길, 김학수, “저사양 기기를 위한 한국어 자동 띄어쓰기 시스템”, *정보처리학회논문지(B)*, 제16권, 제4호, 333-340, 2009.
[3] 심광섭, “CRF를 이용한 한국어 자동 띄어쓰기”, *인지과학*, 제22권, 제2호, 217-233, 2011.
[4] 이창기, 김현기, “Structural SVM 을 이용한 한국어 자동 띄어쓰기”, *한국정보과학회 2012 한국컴퓨터종합학술대회 논문집(B)*, 제39권, 제1호, 270-272, 2012.
[5] 황현선, 이창기, “딥러닝을 이용한 한국어 자동 띄어쓰기”, *한국정보과학회 2016년 한국컴퓨터종합학술대회 논문집*, 738-740, 2016.
[6] 이창기, “사용자가 입력한 띄어쓰기 정보를 이용한 Structural SVM 기반 한국어 띄어쓰기”, *정보과학회논문지: 컴퓨팅의 실제 및 레터*, 제20권, 제5호, 301-305, 2014.
[7] M. Schuster and K. K. Paliwal, Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing, 45.11, pp. 2673-2681, 1997.
[8] 윤보현, 조민정, 임해창, “통계 정보와 선호 규칙을 이용한 한국어 복합 명사의 분해”, *정보과학회논문지(B)*, 제24권, 제8호, 900-909, 1997.