문장 정보를 고려한 딥 러닝 기반 자동 띄어쓰기의 개념 및 활용

조원익[1] , 천성준[1], 김지원[2], 김남수[1]
서울대학교, 전기정보공학부 뉴미디어통신금융연구소[1] 및 언어학과[2]
제 30회 한글 및 한국어 정보처리 학술대회 논문집 (2018년)

Cho, W. I., Cheon, S. J., Kang, W. H., Kim, J. W., & Kim, N. S. (2018). Real-time automatic word segmentation for user-generated text. *arXiv preprint arXiv:1810.13113*.
https://bit.ly/3tChjLu

목차

I. Introduction

- 힌디어, 아랍어, 중국어 등에서 띄어쓰기(Token Segmentation) 관련 연구가 활발히 진행되어옴.
- 한국어에서는 띄어쓰기가 필수임.

  ex. 아버지친구분당선되셨어 -> 아버지/친구분/당선/되셨어 or 아버지/친구/분당선/되셨어
- 요즘에는 모바일기기로 입력하기 때문에 귀찮음, 모호함 또는 시간이 걸림 등의 이유로 제대로 된 띄어쓰기가 입력되지 않고 있음.

➢ **입력할 때 실시간으로 자동 띄어쓰기를 지원하는 보조기술 필요**

➢ **이 시스템은 문법적으로 적합한 것을 제안하는 기존의 수정도구와는 차이가 있다.**

II. Proposed System

대화체 문장 등 정형적이지 않은 문장에 대해 적절한 Token Segmentation (띄어쓰기) 를 지원하는 아키텍처

1) 교육(training)에 사용되는 대화체(비정형) 말뭉치
2) 딥러닝 기반 자동 분할 모듈
3) 다양한 SNS 에 적용할 수 있도록 웹 인터페이스 제시

II. Proposed System

2.1 Corpus

- 한국 드라마 대본 데이터
- 특수문자와 지문(ex. (소리치며), (게걸스럽게 먹으며) 등) 삭제
- 2,000,000만 개의 발화 포함 및 다양한 주제를 포함함.
- 등장 캐릭터 수는 2500여명으로 다양한 인물의 어투, 다양한 단어 등 수집 가능

기존 말뭉치와 본 연구에서 사용한 말뭉치의 차이점

① **비표준성:** 서울 표준어와 맞지 않는 대화체 포함. (방언, 중세 한국어, 외국어번역 등 포함)

② **비정규성:** 문장의 상당부분 정규화 되지 않은 표현 포함 (속삭임, 신조어, 은어, 불완전한 문장 등)

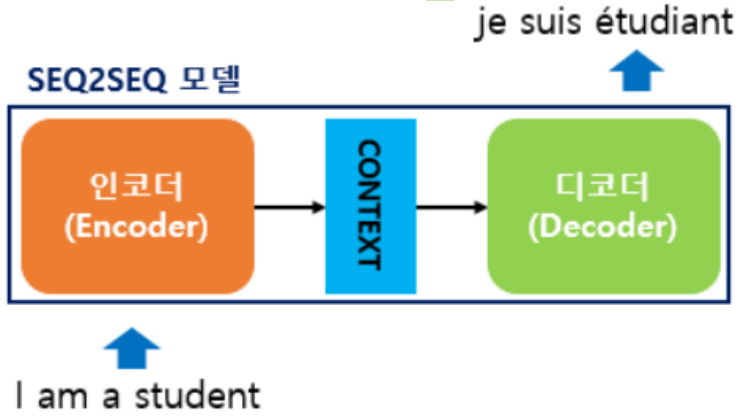
③ **약한 띄어쓰기 규칙:** 문어체처럼 표준 우리말을 엄격히 따르지 않고, 문장 운율에서 말을 할 때 어색함을 유도하지 않는 방식으로 띄어쓰기 적용. ex. 나 너 본 지 한 세 달 다 돼 가 -> 나 너 본지 한 세달 다 돼 가

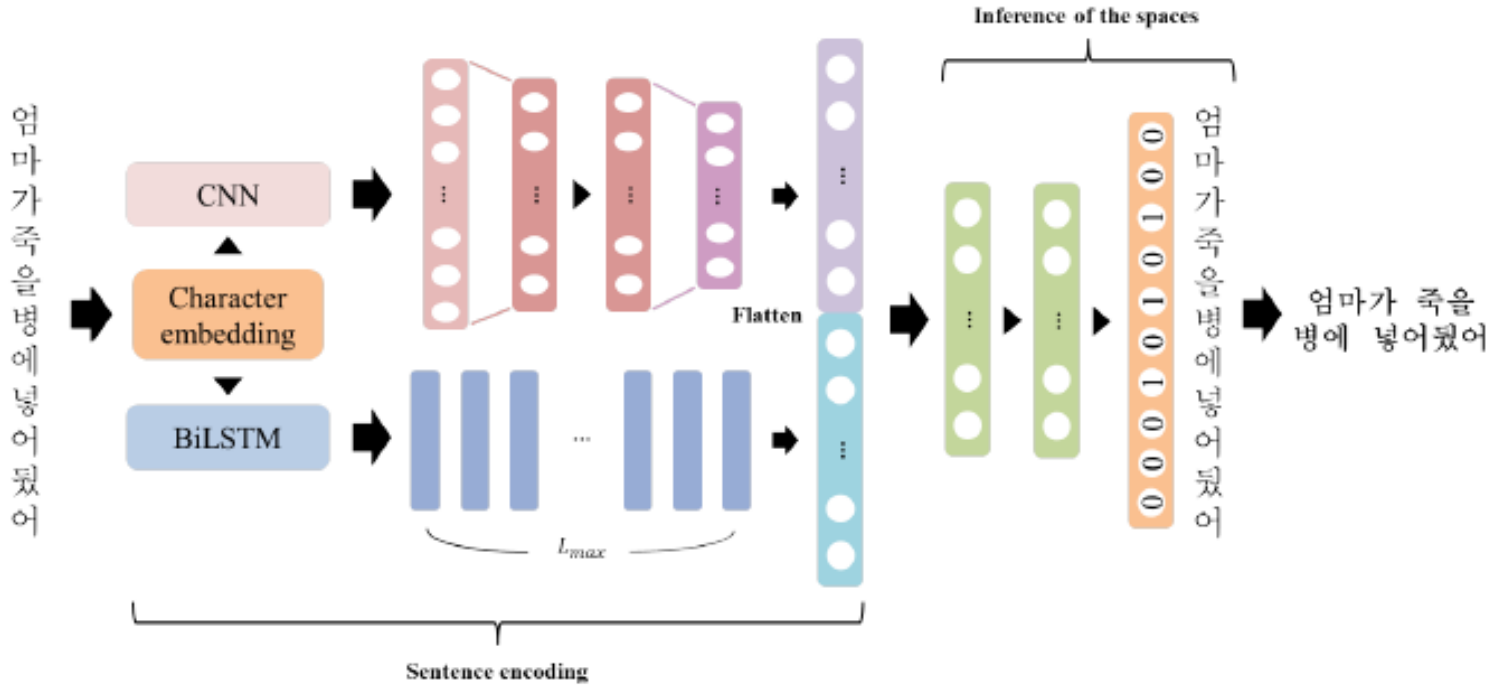
④ **오류 허용:** 발음이 문장의 의미에 영향을 미치지 않는 경우, 오류 허용함. (너 힘들것가타서, 힘들떼 웃는 자가 일류)

III. Proposed System

2.2 Automatic Segmentation Module

Encoder-Decoder 란?





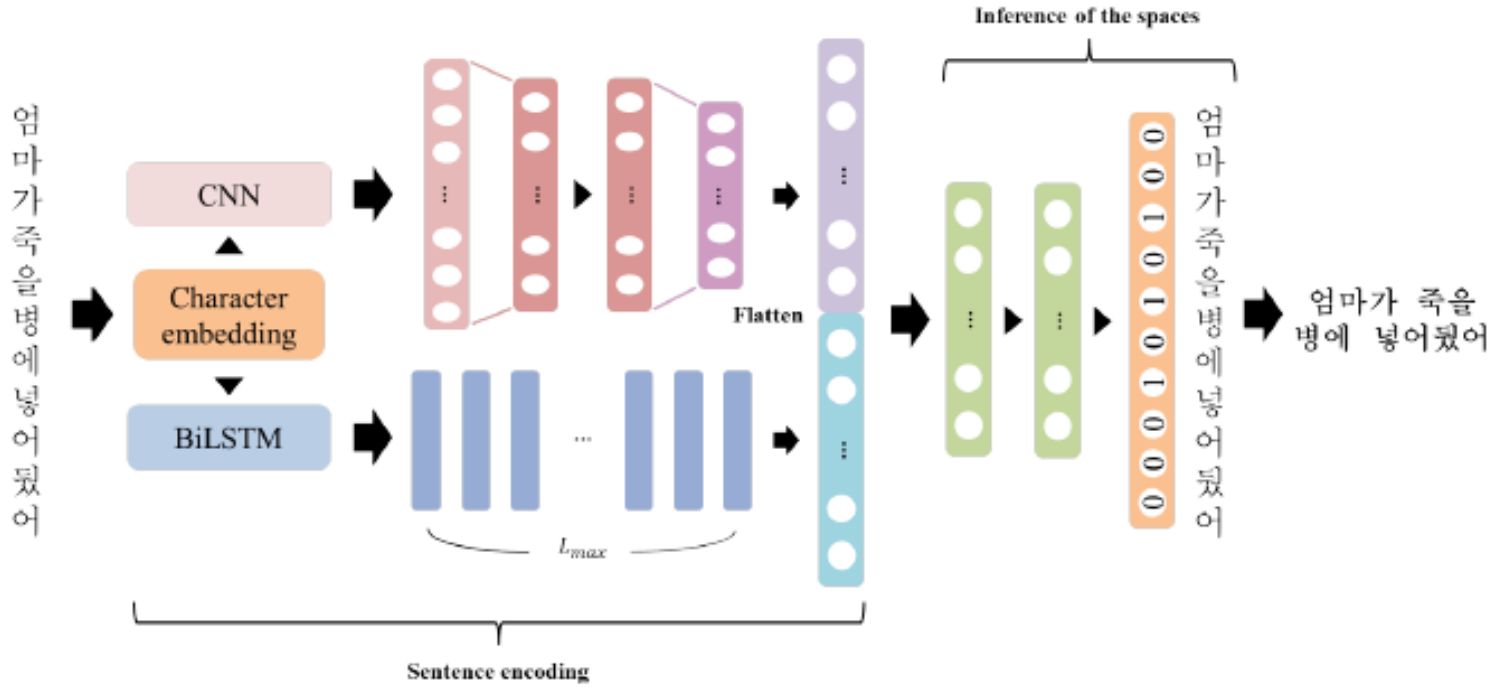
Encoder

```
안녕하세요. 반갑습니다.
['_안녕', '하세요', '.', '_반갑', '습니다', '.']
```

- 문자 임베딩: fasttext의 subword 임베딩을 사용하여 Corpus 훈련, 대상 문장의 각 음절들을 100차원 벡터로 임베딩, 이를 고정된 길이의 vector sequence로 패딩하여 CNN과 BiLSTM의 입력으로 사용
- 입력 문장은 띄어쓰기가 없는 상태(non-segmented sequence)로 제공됨.
- 우리 시스템은 한 문장에서 이웃 단어만을 고려해서 띄어쓰기를 하지 않고, 전체 구성요소를 고려해서 분할한다.

III. Proposed System

2.2 Automatic Segmentation Module



Decoder

- 각 문자 바로 뒤에 공백(띄어쓰기)이 있어야 하는지 {0,1} 이진 시퀀스 유추
- 출력은 L-1이어야 함. (L은 입력문장길이)

ex. 입력문장 11자, 엄마가죽을병에넣어뒀어 -> 출력물 10, [0,0,1, 0, 1, 0, 1, 0,0,0] (마지막 글자 뒤에는 띄어쓰기X)
엄마가V죽을V병에V넣어뒀어

III. Experiment

3. 1 Implementation

- fast-text와 keras 를 사용하여 구현함.
- 최대 input 길이: 100 (발화의 일반적인 길이 고려)
- Input: 분할이 되지 않은 문장에서 추출됨
- Output: Segment 를 나타내는 이진벡터
- Activation Function: ReLU
  (Sigmoid 사용시 과적합이 되어 ReLU 사용)
- Loss Function: MSE

	Specification	
Corpus	Instance of utterances	2,000,000 (#char: >2,500)
CNN	Input size (single channel)	(L_{max}, 100, 1)
	Filters	32
	Window	Conv layer: (3, 100) Max pooling: (2,1)
	# Conv layer	2
BiLSTM	Input size	(L_{max}, 100)
	Hidden layer nodes	32
MLP	Hidden layer nodes	128
	# Layers	2
Others	Optimizer	Adam (0.0005)
	Batch size	128
	Dropout	0.3 (for MLPs)
	Activation	ReLU (MLPs, CNN, output)
	Loss function	Mean squared error (MSE)

Table 1: System specification.

III. Experiment

3. 2 Evaluation

- 본 연구에서 제안한 시스템은 비정형 언어 데이터(대화체, 방언 등)에 대한 실시간 띄어쓰기 지원 기술
- 적합한 데이터셋을 찾을 수 없었고, Accuracy, F1-score로 판단하기에는 무리가 있음. (사람마다 기준이 다름.)
- 따라서 양적 및 질적 평가 수행.
- 비교 모델로는 Naver 맞춤법 검사기, KoSpacing, 부산대 맞춤법 검사기 사용

III. Experiment

- 제안된 시스템은 엄격하게 정확한 세분화를 목표로 하지 않음.
- 복수 정답이 포함된 결과를 평가하기 위해 평균의견점수(MOS)를 수정하여 사용.
- 각 시스템의 최종 순위 R은 전체 입력 발언에 대한 순위의 평균과 같음

랭킹 스코어 S, 시스템 수 N=3, R=랭킹

$$S = \frac{N + 1 - R}{N}$$

ex. 1위 $\frac{3+1-1}{3} = 1$ 2위 $\frac{3+1-2}{3} = \frac{2}{3}$ 3위 $\frac{3+1-3}{3} = \frac{1}{3}$

- S의 평균을 각 모델의 Score로 사용
- 14명의 한국인에게 순위 배정요구. 둘 이상의 동일한 순위 허용 (1,1,2), (1,2,2) 등
- 박막례님의 인스타그램 포스트 30개를 대상으로 평가(띄어쓰기X, 맞춤법X 등)

Naver: 0.6809, KoSpacing: 0.7142, Proposed: 0.7444

애순이는아주좋은친구다백연친구
다참좋은친구다감동바덧다

...

인생잘산거야힘들떼같이걱정하는
친구가그게친구야애순아나우리는
백연친구야

....

Cho, W. I., Cheon, S. J., Kim, J. W., & Kim, N. S. (2018). Real-time Automatic Word Segmentation for User-generated Text. arXiv

IV. Experiment

나너본지한세달다돼감

- Naver: 나/너/본/지/한/세/달/다/돼/감
- Proposed: 나/너/본지/한/세달/다/돼/감

➢ 네이버는 띄어쓰기 엄격하게 준수, 하지만 제안된 모델은 구문적으로 가까운 음절을 클러스터링하여 운율적 지속시간을 더 자연스럽게 함. 이는 드라마 대본의 특성에 영향을 받은 결과로 보임.

아버지친구분당선되셨어요

- KoSpacing: 아버지/친구/분당선/되셨어요
- Proposed: 아버지/친구분/당선/되셨어요

➢ KoSpacing은 구어적 표현보다 교통관련기사로 훈련되었기 때문에 위와 같이 Segmentation을 한 것으로 보임.

뭣이중헌지도모름서

- Naver: 뭣이중헌지도모름서
- Proposed: 뭣이/중헌지도/모름서

➢ Naver 는 방언에 대해 대응하지 않음. proposed 모델은 누락이나 변형없이 방언에도 적절한 결과를 보여줌.

Cho, W. I., Cheon, S. J., Kim, J. W., & Kim, N. S. (2018). Real-time Automatic Word Segmentation for User-generated Text. arXiv

IV. Experiment

표 1. 다양한 문장들을 통한 Web-based 맞춤법 검사기들과의 비교

입력 문장	*부산대*	*Naver*	*Proposed*
1. 그것도안알아봤냐	그것도 안 알아봤냐	그것도 안 알아봤냐	그것도 안 알아봤냐
2. 왜말을못해왜말을못하냐구	왜 말을 못해 왜 말을 못하냐고	왜 말을 못 해 왜 말을 못 하냐고	왜 말을 못해 왜 말을 못하냐구
3. 아버지친구분당선되셨더라	아버지 친구분 당선되셨더라	아버지 친구 분당서 되셨더라	아버지 친구분 당선 되셨더라
4. 엄마가죽을병에넣어뒀어	엄마가 죽을병에 넣어뒀어	엄마가 죽을 병에 넣어뒀어	엄마가 죽을 병에 넣어뒀어
5. 나열만큼사랑해	[대치어 없음]	나열 만큼 사랑해	나 열만큼 사랑해
6. 상하이스파이스치킨콤보부탁해	상하이 스파이스치킨 캄보 부탁해	상하이 스파이스 치킨 콤보 부탁해	상하이 스파이스 치킨콤 보부탁해
7. 인싸연구자가되는방법은	인사연구자가 되는 방법은	인사 연구자가 되는 방법은	인싸 연구자가 되는 방법은
8. 네이버와부산대의맞춤법검사	네 이번과 부산대의 맞춤법검사	네이버와 부산대의 맞춤법검사	네 이버와 부산대의 맞춤 법 검사
9. 그는눈을질끈감았다	그는 눈을 질끈 감았다	그는 눈을 질끈 감았다	그는 눈을 질끈감았다
10. 이함무봐라	[대치어 없음]	이함 무 봐라	이 함 무봐라
11. 정신똑띠차리라	정신 똑똑히 차리라	정신똑띠차리라	정신 똑띠 차리라
12. 뭣이중헌지도모름서	뭣이 중한지도 모르면서	뭣이중헌지도모름서	뭣이 중헌지도 모름서

1~2: 일상 대화에서 다른 모듈들이 문장을 수정하는 것과 달리 본 시스템은 띄어쓰기만 보조함.

3~5: 어색함을 유발할 수 있는 띄어쓰기(혹은 띄어 쓰지 않기)를 제시한 시스템은 잘 회피함.

5번: 다른 모듈들이 응답을 출력하지 않거나, 내용을 왜곡한데 반하여 제시한 시스템은 비표준 구어를 잘 유지

6~8: 제시한 시스템이 외래어, 신조어, 최신 고유명사에 취약한 모습을 보임. => 페이스북, 인스타그램 등 다양한 커뮤니티의 게시글 코퍼스 학습을 통해 해결할 수 있을 것으로 보임.

10~12: 비표준어를 입력하였음에도 불구하고 띄어쓰기를 바람직하게 수행함.

Cho, W. I., Cheon, S. J., Kim, J. W., & Kim, N. S. (2018). 문장 정보를 고려한 딥러닝 기반 자동 띄어쓰기의 개념 및 활용. 한글 및 한국어 정보처리학회 2018

V. Conclusion

우리 연구의 구별되는 특징은 다음과 같다.

- 한국어 대본의 특성과 현실세계의 요구에 기인하는 Motivation
- noisy 한 사용자 생성 한글 텍스트의 실시간 자동 세분화에 적합한 Corpus 사용 및 모델 아키텍처 제안
- 기존 도구와의 정성적인 비교 연구, 특히 샘플 스크립트를 활용한 양적 및 질적 평가 수행

✓ 비정형 한국어 데이터에 사용 가능한 Token Segmentation 시스템 제안

✓ 이 시스템은 디지털 소외계층을 위한 보조기술 및 노이즈가 많은 텍스트 분석을 위한 시스템으로 활용될 수 있음.

✓ 나아가 스마트 장치의 자동 음성인식, TTS(Text-to-Speech) 등의 사전처리 시스템으로 활용 가능

Cho, W. I., Cheon, S. J., Kim, J. W., & Kim, N. S. (2018). 문장 정보를 고려한 딥러닝 기반 자동 띄어쓰기의 개념 및 활용. 한글 및 한국어 정보처리학회 2018

코드 실습

Google Colab(띄쓰봇, Okt(), Kkma(), KoSpacing)
https://colab.research.google.com/drive/1T2fllDks0IY9q1XEHQGBG2sVwb8_Zhnx#scrollTo=28GJf8JLEIQs

Konlpy (Okt, Kkma 등)
https://konlpy.org/ko/latest/index.html

KoSpacing
https://github.com/haven-jeon/PyKoSpacing