딥러닝 기반 한국어 띄어쓰기 모델

Korean Word-Spacing

신재영 연구원

INDEX

1. Introduction

- 한국어는 띄어쓰기에 따라 문장의 의미와 뜻이 달라질 수 있어 띄어쓰기 교정이 매우 중요
  ex) 아버지가방에들어가신다
- AI 챗봇 학습 데이터 구축 과정에서 띄어쓰기가 중요
  ex) 대우조선해양 PINBOT 프로젝트
- 최근 자연어처리에서 딥러닝 모델 및 라이브러리에 대한 연구가 활발히 진행 중
  ex) 한글 및 한국어 정보처리 학술대회 논문집
- 딥러닝 기반의 한국어 띄어쓰기 주요 모델을 소개하고 비교 분석하여 자연어처리 관련 프로젝트에 적용하고자 함

1. Introduction

- **음절 vs 어절**

- 음절 : 한 번에 소리 낼 수 있는 소리마디이며 의미와 전혀 관계가 없는 음성학적 문법 단위
- 어절 : 한 단어 및 그 이상의 이어진 단어들에 의하여 이루어진 문법단위. 문장단위

음절 vs 어절

2. Word-Spacing Method

띄어쓰기 방법론

- **규칙 기반 방법**

- 언어학적 자원을 요구하며 어휘 지식, 사전정보를 이용한 휴리스틱을 적용하는 방식
- 미등록어에 취약하고, 규칙의 작성 및 유지가 힘들다는 한계
- ex) 형태소 분석

아버지	가	방	에	들어가신다
Noun	Josa	Noun	Josa	Verb

형태소 분석 예시

2. Word-Spacing Method

띄어쓰기 방법론

- **통계 기반 방법**

- 대규모 말뭉치로부터 확률 및 통계 정보를 이용하여 어절의 경계에 공백을 삽입하는 형태
- 미등록어 문제에 대해 어느 정도 해결 가능하나 대량의 데이터가 필요
- ex) CRF

- **CRF(Conditional Random Field)**

- 주어진 입력 데이터 열에 대하여 레이블 열의 확률을 이용하는 조건부 확률 모델
- 입력 데이터 열에 레이블 열을 부여하는 문제에 적합 -> Sequence Labeling Task

  ex) 순차적 레이블링(Sequence Labeling) : 입력 $X = [x_1,\ x_2,\ x_3, \dots,\ x_n]$ 에 대하여 레이블 $y = [y_1,\ y_2,\ y_3, \dots,\ y_n]$ 부여
- BIO 태그를 적용 -> B(Begin), I(Inside), O(Outside)

$$P(y|x) = \frac{exp(\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(x, i, y_i, y_{i-1}))}{\sum_{y'} exp(\sum_{j'=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(x, i, y_i', y_{i-1}'))}$$

나	는	바	보	다
B	I	B	I	I

$$\frac{1}{2^5}$$

2. Word-Spacing Method

띄어쓰기 방법론

- **딥러닝 기반 방법**

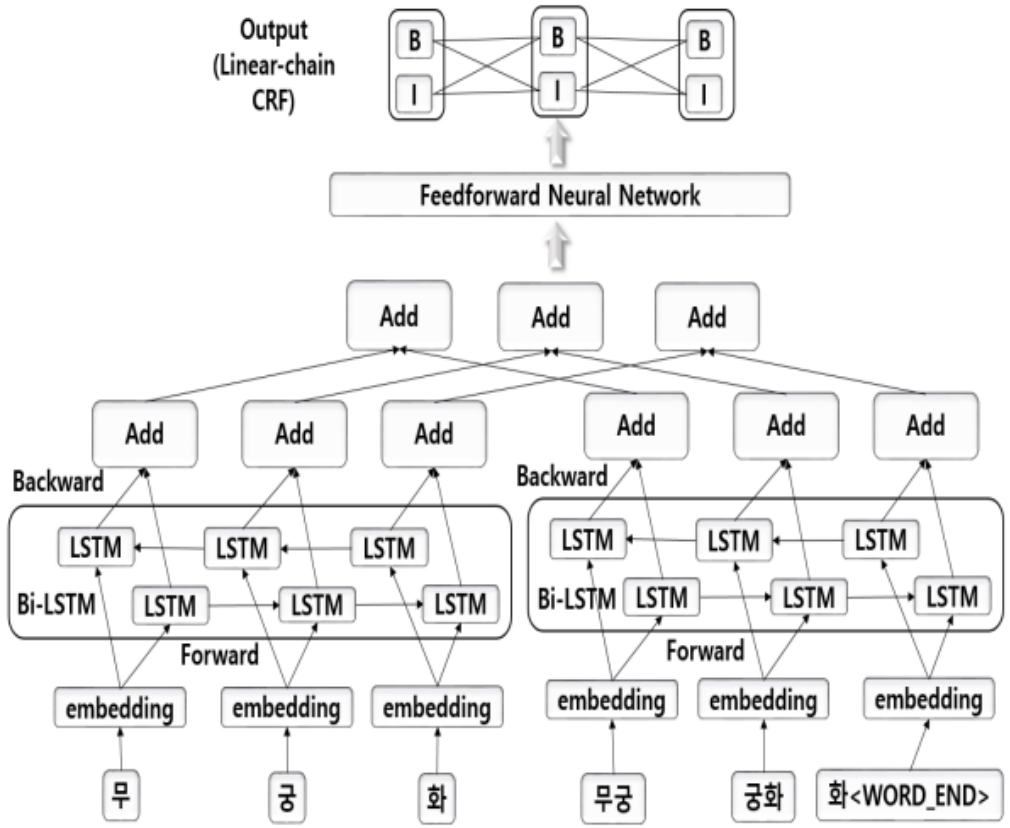
- 통계 기반 방법에 이어 딥러닝 기반의 띄어쓰기 모델 또한 순차적 레이블링 Task로 접근
- 순차적 레이블링에 효과적인 CRF 활용하여 RNN, LSTM, GRU, CNN 등의 다양한 딥러닝 기반 모델 연구 중
- 기존 연구에서 발전된 딥러닝 기반의 한국어 띄어쓰기 모델들을 살펴 보고 분석하고자 함

3. Korean Word-Spacing Model

(1) Bidirectional LSTM-CRF

이현영, 강승식, "음절 임베딩과 양방향 LSTM-CRF를 이용한 한국어 문장 자동 띄어쓰기", 제30회 한글 및 한국어 정보처리 학술대회 논문집, 2018

Model Architecture



음절 Unigram과 Bigram을 이용한 양방향 LSTM-CRF

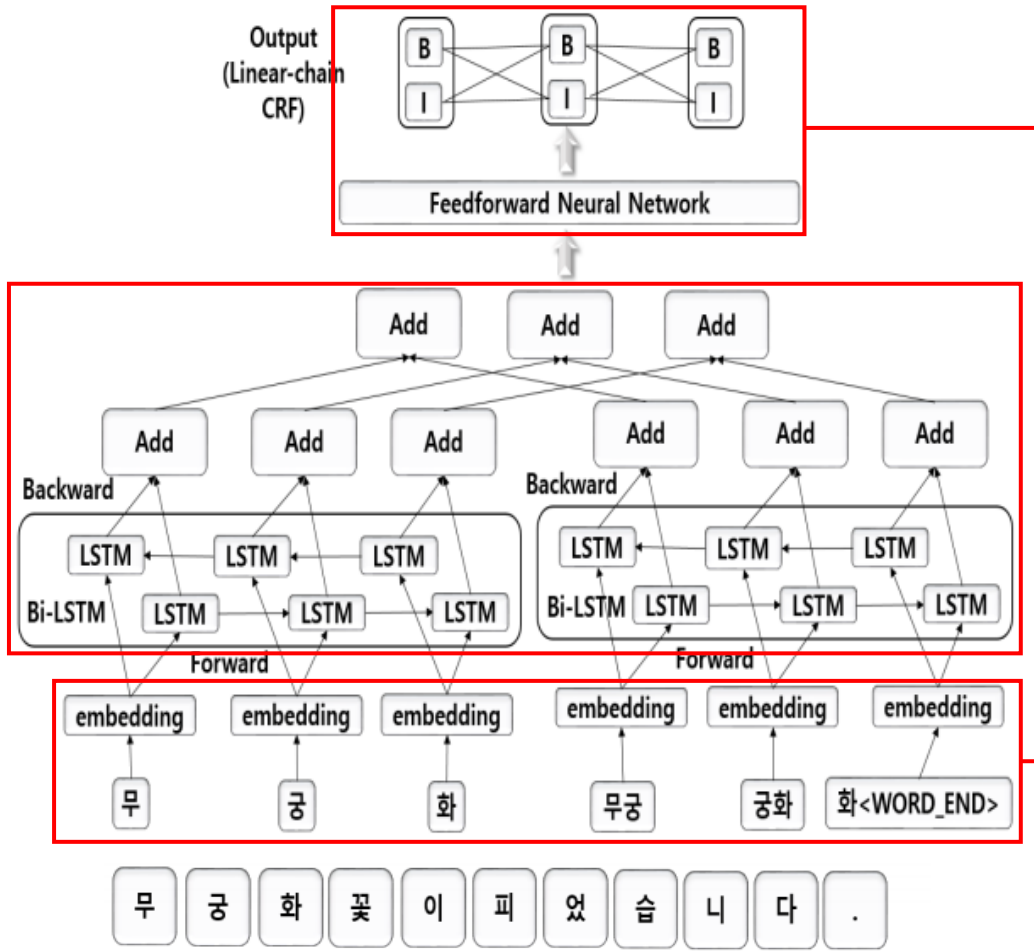
- 띄어쓰기가 전혀 적용되지 않은 문장을 입력으로 입력문장의 각 음절에 해당하는 띄어쓰기 태그 클래스(B 또는 I)로 분류
- 양방향 LSTM-CRF 모델을 한국어 자동 띄어쓰기 문제 적용

3. Korean Word-Spacing Model

(1) Bidirectional LSTM-CRF

이현영, 강승식, "음절 임베딩과 양방향 LSTM-CRF를 이용한 한국어 문장 자동 띄어쓰기", 제30회 한글 및 한국어 정보처리 학술대회 논문집, 2018

Model Architecture



무	궁	화	꽃	이	피	었	습	니	다	.
B	I	I	B	I	B	I	I	I	I	I

CRF를 통한 태그 열 예측

Vocabulary Dictionary :	Uni-gram	Bi-gram
	<PAD>, <UNK> 무, 궁, 화, 꽃, 이, 피, 었, 습, 니, 다, .	무궁, 궁화, 화꽃, 꽃이, 이피, 피었, 었습, 습니, 니다, 다. , .<WORD_END>

말뭉치 기반의 Unigram과 Bigram

3. Korean Word-Spacing Model

(1) Bidirectional LSTM-CRF

이현영, 강승식, "음절 임베딩과 양방향 LSTM-CRF를 이용한 한국어 문장 자동 띄어쓰기", 제30회 한글 및 한국어 정보처리 학술대회 논문집, 2018

Experiments

	Train	Test	Total
문장 수	13,500	1,500	15,000
단어 (중복 단어 포함)	277,718	31,107	308,825
음절 (중복 음절 포함)	882,134	98,774	980,908

차세정 언어처리 경진대회 말뭉치 데이터

Model	Syllable Embedding	Bi-LSTM's Output Operation	Embedding and LSTM cell unit size
1	Unigram	Add	250
2			300
3	Unigram + Bigram	Add	250
4			300

음절 임베딩 종류와 셀 유닛 크기에 따른 모델

3. Korean Word-Spacing Model

(1) Bidirectional LSTM-CRF

이현영, 강승식, "음절 임베딩과 양방향 LSTM-CRF를 이용한 한국어 문장 자동 띄어쓰기", 제30회 한글 및 한국어 정보처리 학술대회 논문집, 2018

Experiments

Model	Spacing Recall	Syllable Accuracy	Word Recall	Word Precision	Word F1 Score
1	93.80	96.346	86.356	86.518	86.437
2	92.77	96.178	84.906	86.319	85.607
3	**95.25**	97.332	**89.526**	**90.032**	**89.779**
4	95.19	**97.337**	89.410	90.030	89.719

성능 평가 표 (단위 : &)

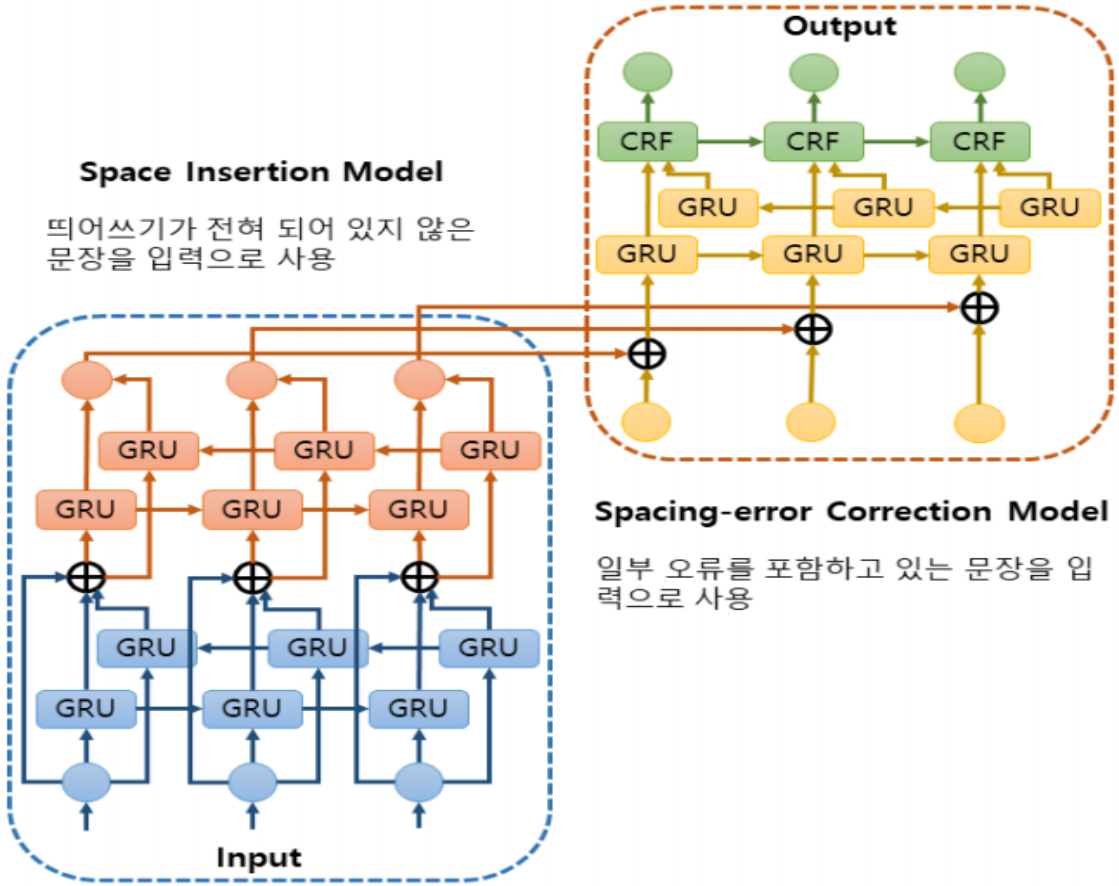
- **평가 척도**
- Spacing Recall : 공백 재현율
- Syllable Accuracy : 띄어쓰기 태그 정확도
- Word Recall : 어절 재현율
- Word Precision : 어절 정확도
- F1 score

3. Korean Word-Spacing Model

(2) Bidirectional GRU-CRFs

최기현, 김시형, 김학수, "심층신경망 기반 2단계 한국어 자동 띄어쓰기 모델", 제30회 한글 및 한국어 정보처리 학술대회 논문집, 2018

Model Architecture



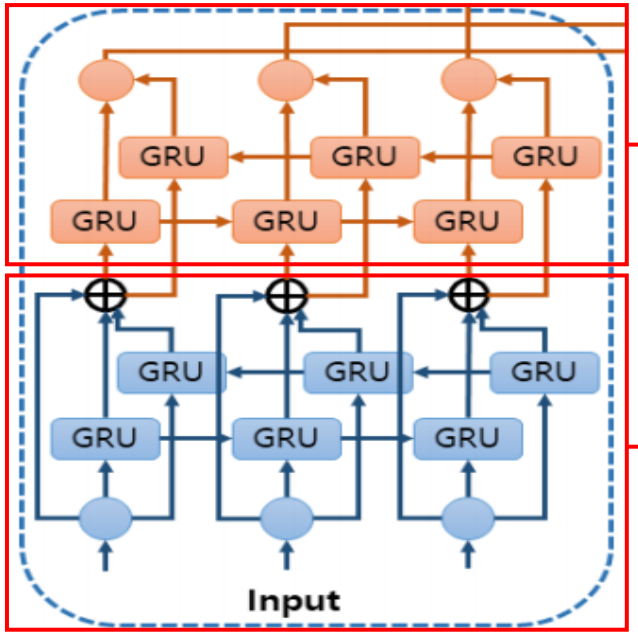
- 기존의 연구들은 주로 문장의 공백을 모두 제거한 문장을 입력으로 사용하는 것에 집중
- 입력 문장의 띄어쓰기 정보를 모두 생략한다면 사용자가 올바르게 입력한 띄어쓰기 정보까지 삭제될 수 있음
- BiGRU-CRFs를 기반으로 띄어쓰기가 안된 문장을 자동으로 띄어쓰기 해주고, 사용자의 띄어쓰기 정보를 이용하여 오류를 교정해줌으로써 성능을 향상 시키는 모델 제안

3. Korean Word-Spacing Model

(2) Bidirectional GRU-CRFs

최기현, 김시형, 김학수, "심층신경망 기반 2단계 한국어 자동 띄어쓰기 모델", 제30회 한글 및 한국어 정보처리 학술대회 논문집, 2018

Model Architecture



Space Insertion Model

2. 띄어쓰기가 된 문장을 출력

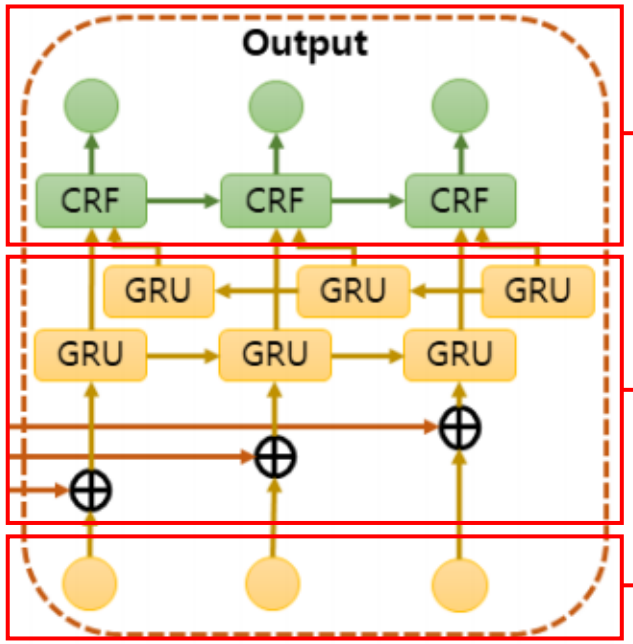
1. 문장을 입력 받고 임베딩 벡터와 각 음절에 대한 입력 정보(명사 사전) 사용

3. Korean Word-Spacing Model

(2) Bidirectional GRU-CRFs

최기현, 김시형, 김학수, "심층신경망 기반 2단계 한국어 자동 띄어쓰기 모델", 제30회 한글 및 한국어 정보처리 학술대회 논문집, 2018

Model Architecture



Spacing-error Correction Model

5. CRF를 통해 최종적으로 각 음절에 태그(B, I) 부착

4. 삽입 모델의 출력 문장과 오류교정 모델 입력 문장을 연결

3. 띄어쓰기가 된 문장에 임의로 오류를 발생시킨 정보를 입력으로 사용

3. Korean Word-Spacing Model

(2) Bidirectional GRU-CRFs 최기현, 김시형, 김학수, "심층신경망 기반 2단계 한국어 자동 띄어쓰기 모델", 제30회 한글 및 한국어 정보처리 학술대회 논문집, 2018

Experiments

	뉴스 말뭉치	세종 말뭉치	ETRI 품사 태그 말뭉치
Train	253,138 문장	257,071 문장	-
Test	-	-	27,854 문장

훈련 및 평가 데이터

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

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

에러율 30% 예시

3. Korean Word-Spacing Model

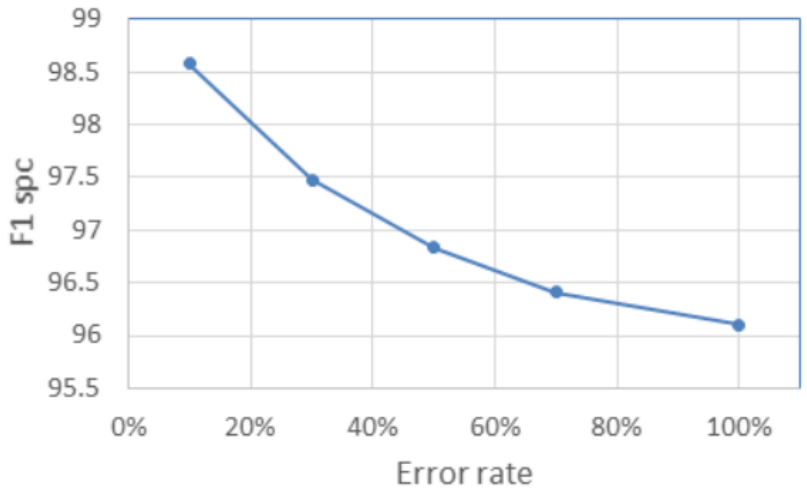
(2) Bidirectional GRU-CRFs

최기현, 김시형, 김학수, "심층신경망 기반 2단계 한국어 자동 띄어쓰기 모델", 제30회 한글 및 한국어 정보처리 학술대회 논문집, 2018

Experiments

	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

에러율에 따른 모델 성능 평가 표 (단위 : &)



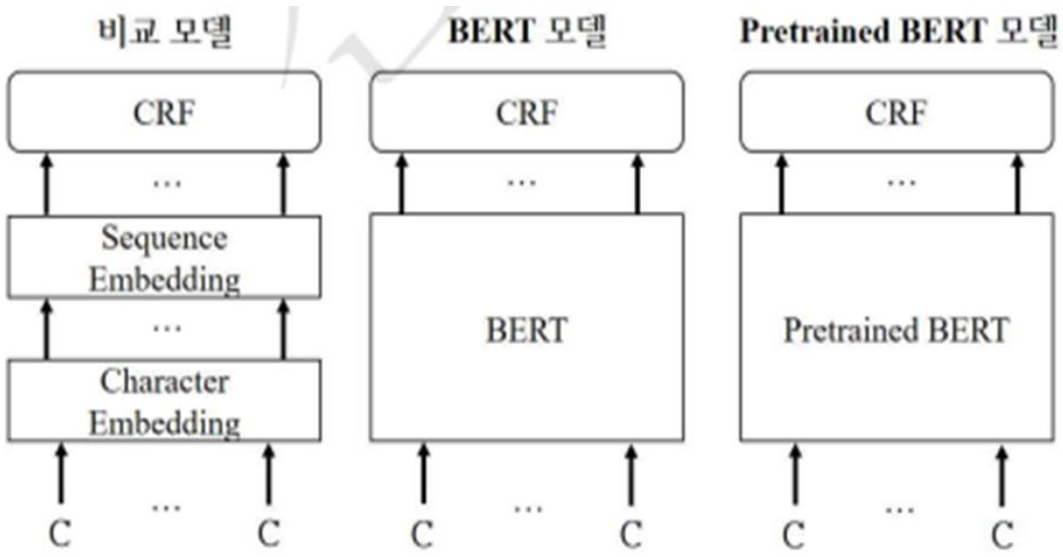
에러율에 따른 모델 성능 변화 그래프

3. Korean Word-Spacing Model

(3) Pretrained BERT-CRF

황태욱, 정상근, "BERT를 이용한 한국어 자동 띄어쓰기", 한국소프트웨어종합학술대회 논문집, 2019

Model Architecture



전체 모델 구조도

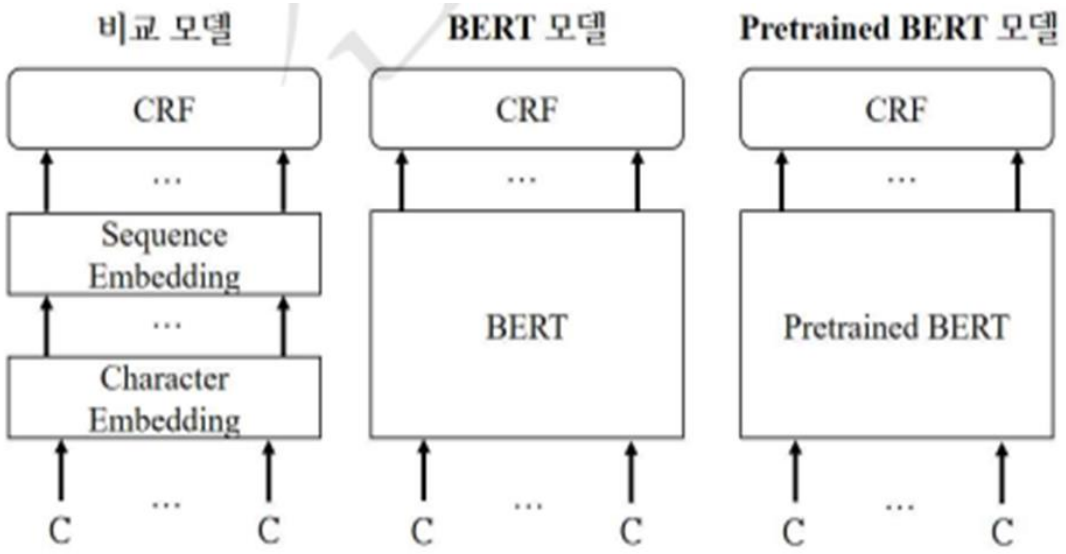
- 자연어처리 분야에서 사전학습 된 BERT와 XLNet 등을 활용한 모델들이 높은 성능을 보여주고 있음
- BERT 모델은 적은 양의 최적 학습만으로도 큰 성능 향상이 검증됨
- 사전학습 된 BERT 모델이 한국어 띄어쓰기에 강인함을 보이고자 함

3. Korean Word-Spacing Model

(3) Pretrained BERT-CRF

황태욱, 정상근, "BERT를 이용한 한국어 자동 띄어쓰기", 한국소프트웨어종합학술대회 논문집, 2019

Model Architecture



전체 모델 구조도

- **사용 모델**
- 비교 모델 : 단방향 LSTM-CRF
- BERT 모델 : 사전학습 되지 않은 BERT-CRF 모델
- Pretrained BERT 모델 : 사전학습 된 BERT-CRF 모델
- CRF를 통해 최종적으로 각 음절에 태그(B, I)를 부착하는 방식
- 사전학습 된 BERT 모델과 사전학습 되지 않은 모델의 성능 차이 확인하고 기존의 LSTM-CRF와 비교

3. Korean Word-Spacing Model

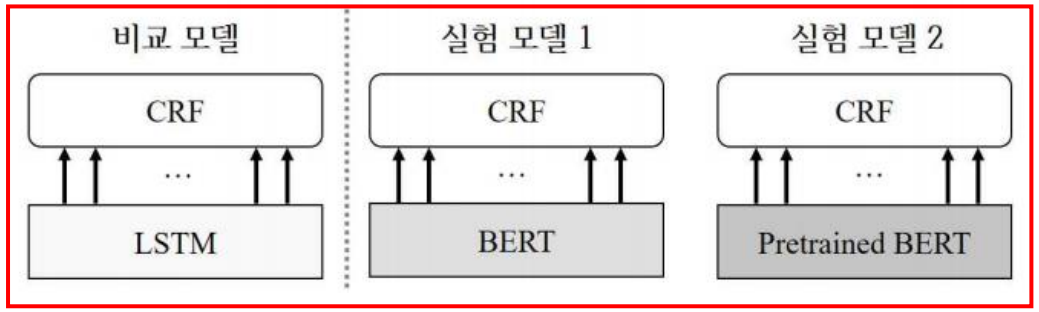
(3) Pretrained BERT-CRF

황태욱, 정상근, "BERT를 이용한 한국어 자동 띄어쓰기", 한국소프트웨어종합학술대회 논문집, 2019

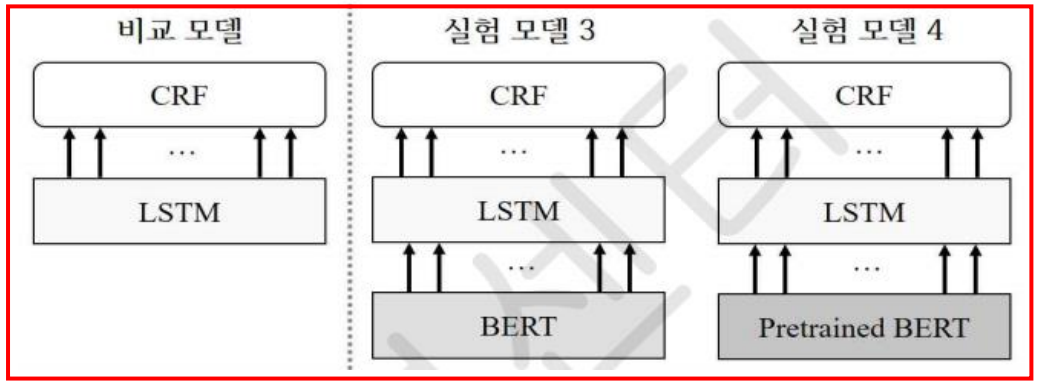
Experiments

	세종 말뭉치
Train	805,112 문장
Test	50,000 문장

훈련 및 평가 데이터



실험군 1



실험군 2

3. Korean Word-Spacing Model

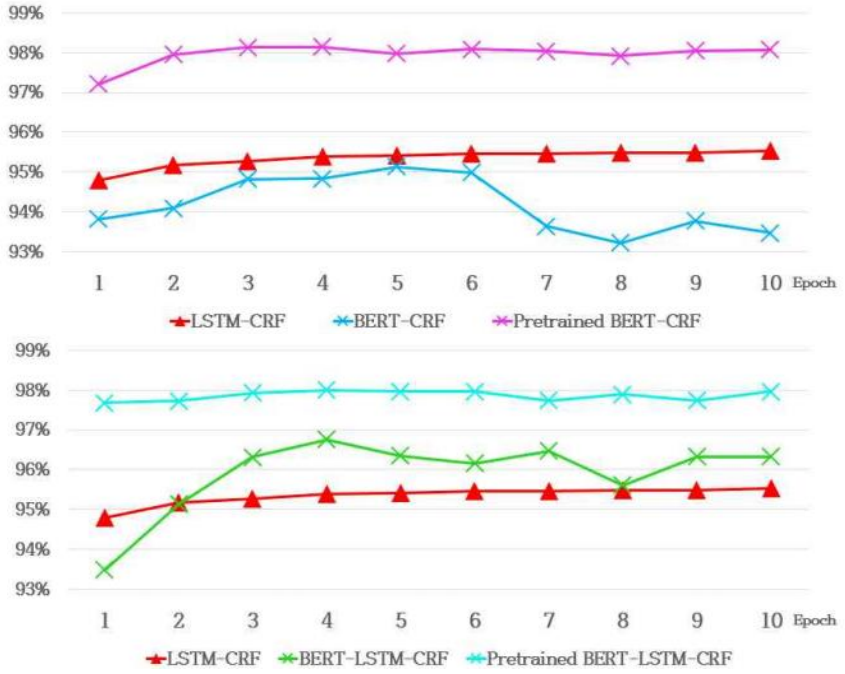
(3) Pretrained BERT-CRF

황태욱, 정상근, "BERT를 이용한 한국어 자동 띄어쓰기", 한국소프트웨어종합학술대회 논문집, 2019

Experiments

방법		음절 단위 정확도
CRFs [8]		97.99%
BiLSTM-CRF + 자질 정보 [2]		97.04%
GRU-CRF + 자질 정보 [1]		**98.32%**
비교 모델	LSTM-CRF	95.53%
실험 모델 1	BERT-CRF	95.13%
실험 모델 2	Pretrained BERT-CRF	**98.14%**
실험 모델 3	BERT-LSTM-CRF	96.75%
실험 모델 4	Pretrained BERT-LSTM-CRF	97.99%

실험 모델 성능 평가 표



Epoch에 따른 성능 변화 그래프

4. Conclusion

(1) 결과 비교

	Bidirectional LSTM-CRF	Bidirectional GRU-CRFs	Pretrained BERT-CRF
학습 및 평가데이터	차세정 언어처리 경진대회 말뭉치	뉴스 말뭉치, 세종 말뭉치, ETRI	세종 말뭉치
데이터 총 개수	1,500 문장	538,063 문장	855,112 문장
입력 정보	Unigram + Bigram	명사 사전	X
모델 성능	89.78%	96.1%	98.14%

4. Conclusion

(2) 최종 결론

- 각 모델 비교 결과 Pretrained BERT-CRF 모델은 입력 정보 없이도 가장 우수한 성능을 보임
- Pretrained BERT 모델은 1~2회의 epoch 만으로 높은 성능을 보여줌
- Pretrained BERT-CRF 모델에 입력 정보를 적용하면 성능이 더 높을 것으로 예상
- 차후 자연어처리 관련 프로젝트에 Pretrained BERT-CRF 모델을 구현해보고 적용하고자 함

Q&A