한국어 ELMo 모델을 이용한 의미역 결정

박찬민, 박영준
Sogang_Alzzam

Naver NLP Challenge

서강대학교
자연어처리 연구실

목차

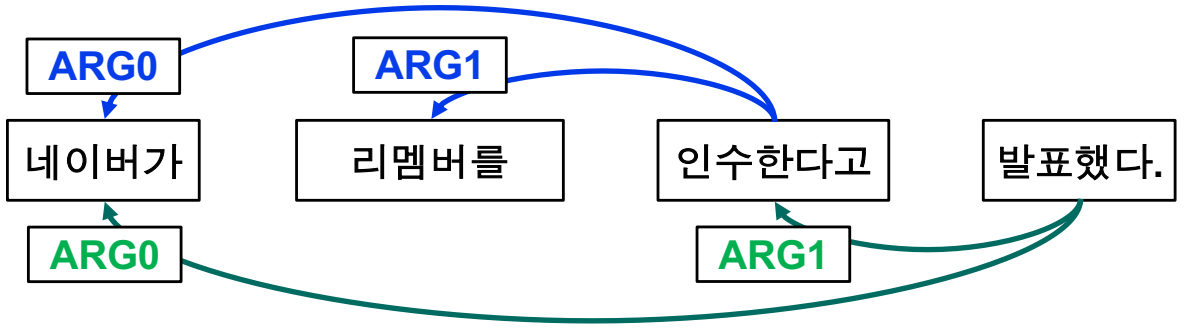
- 서론
- 제안모델
- 실험
- 결론

서론

의미역 결정이란?

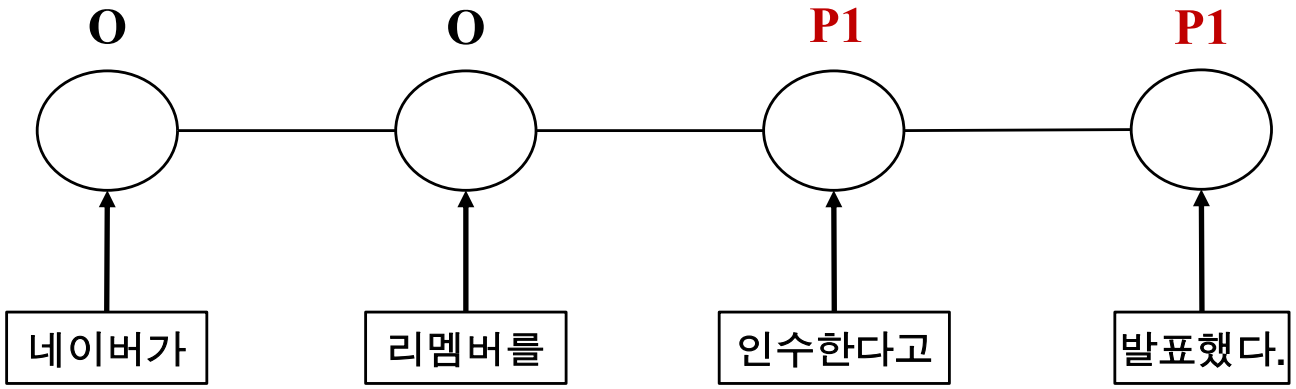
- 문장의 술어를 찾고, 그 술어와 연관된 논항들 사이의 의미관계를 결정하는 문제
  - 논항 : 의미역이 부여된 각 명사구
  - 의미역 : 술어에 대한 명사구의 의미 역할

- "누가, 무엇을, 어떻게, 왜" 등의 의미 관계를 찾아내는 작업

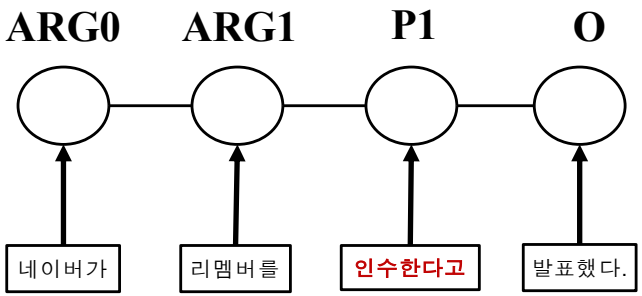


서론

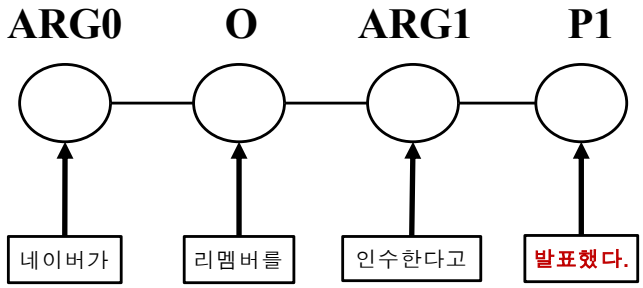
- 의미역 결정 문제를 순차 레이블링 문제로 간주
- **Step1)** 서술어 인식/분류



- **Step2)** 논항 인식/분류
  - "인수한다고"의 논항

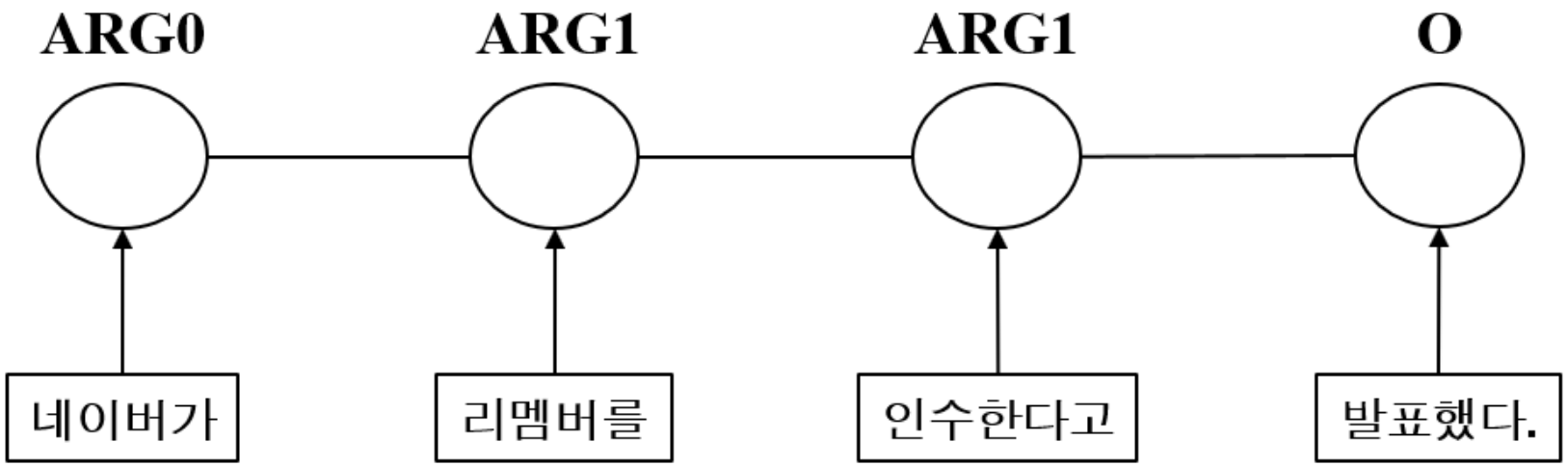


  - "발표했다"의 논항



서론

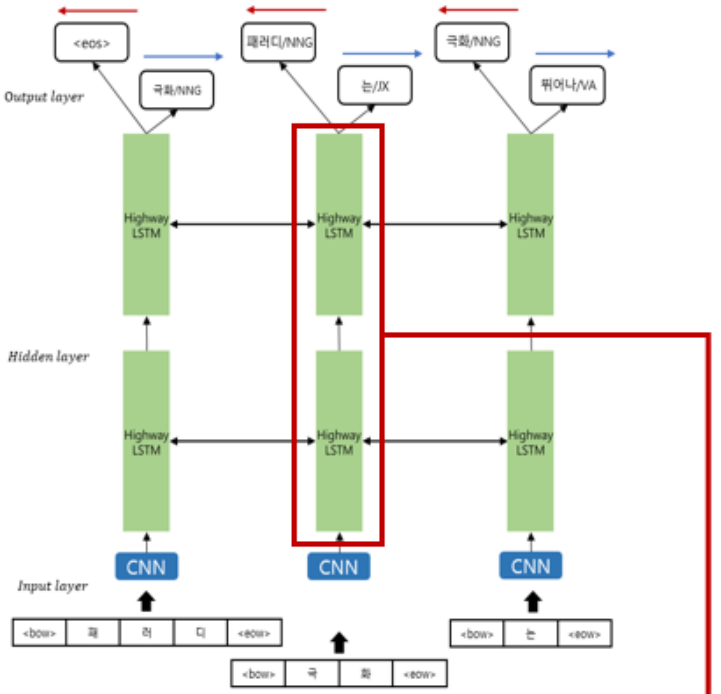
- 입력 문장 전체에 대한 논항 인식/분류 모델 사용

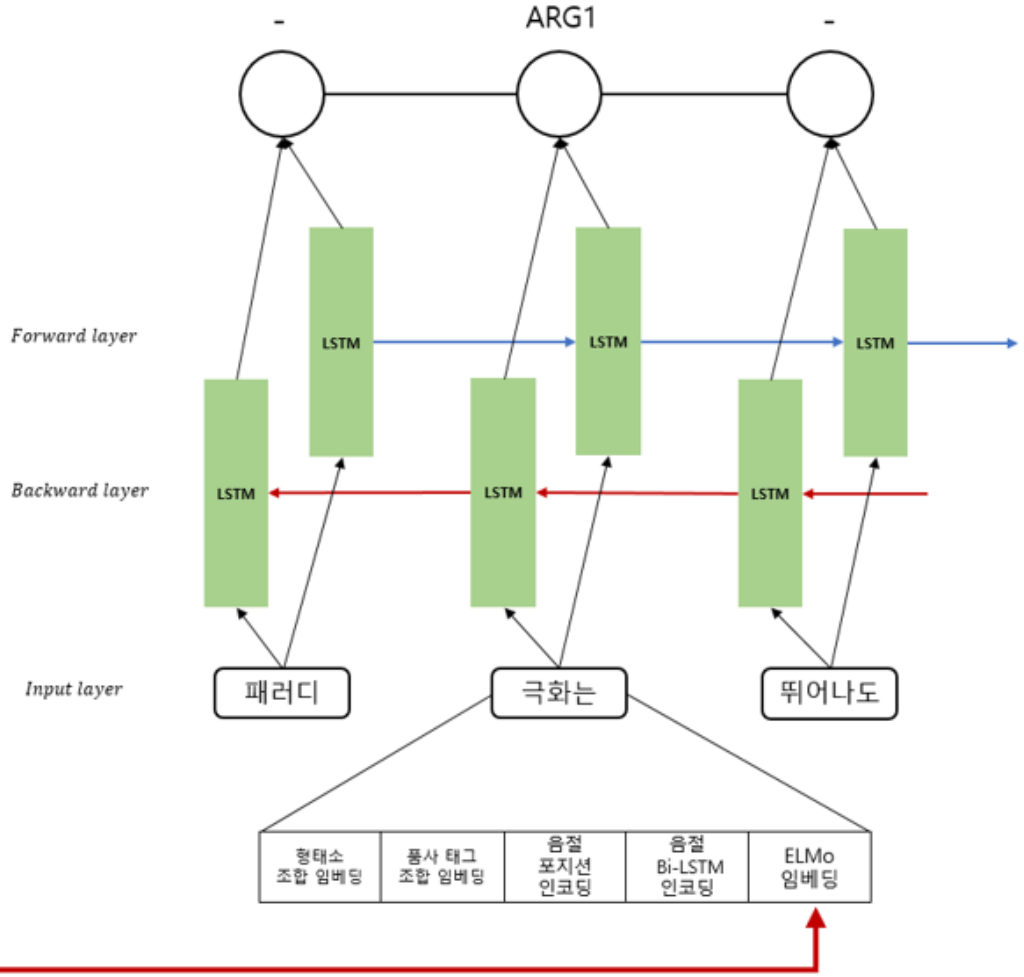


제안 모델

제안 모델

- Bi-LSTM-CRFs
- ELMo



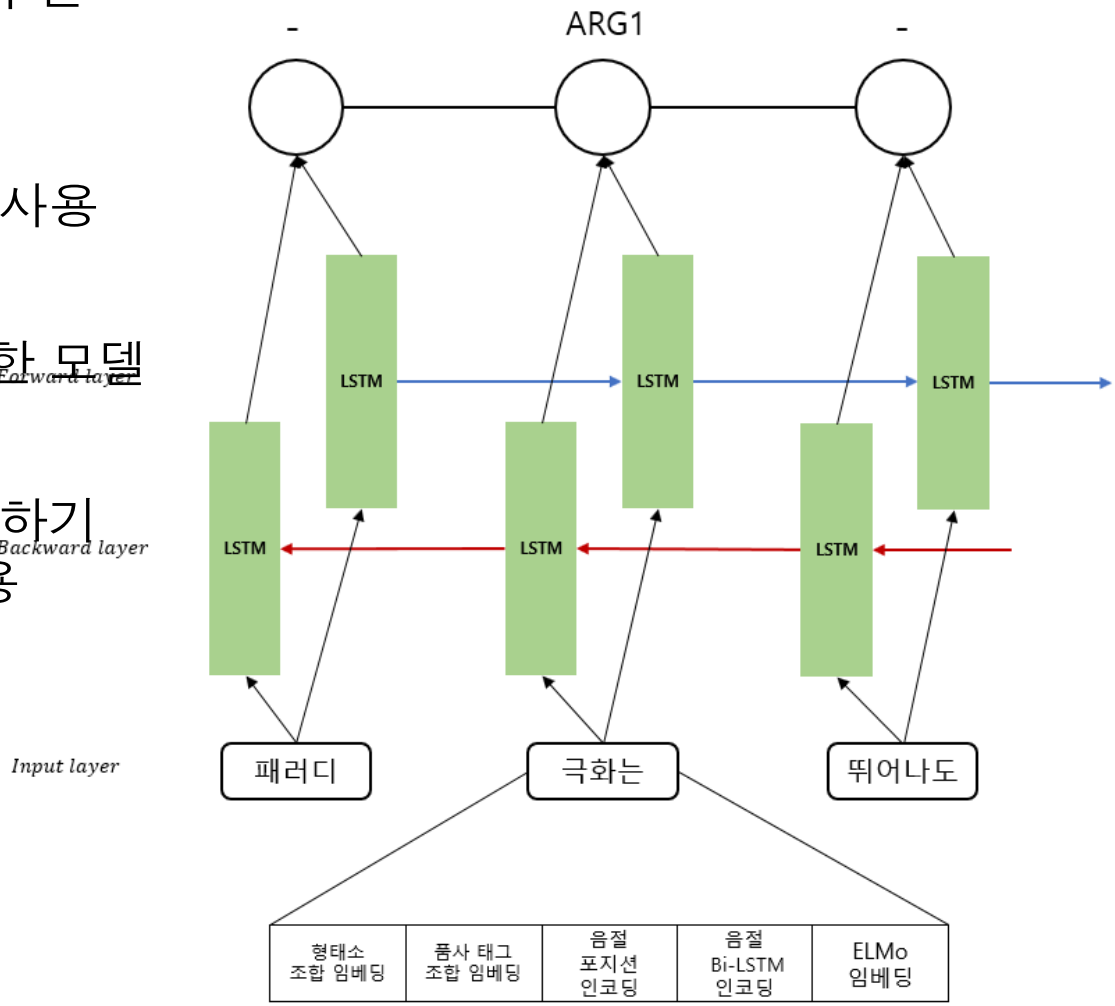


패러디 극화는 뛰어나도 그 원작을 본 사람들은 수긍이 어렵다는 게 장점이다.

제안 모델

Bi-LSTM-CRFs

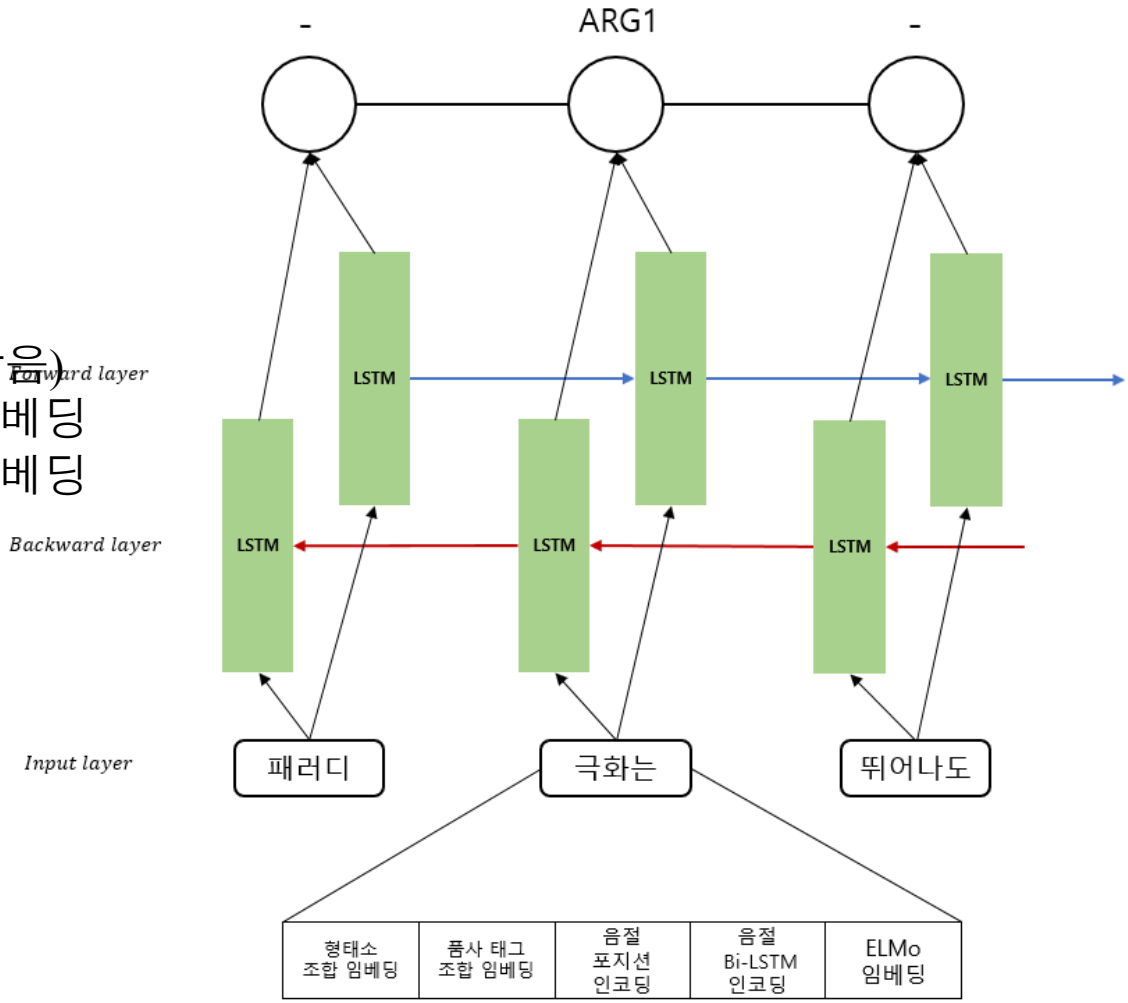
- 순차 레이블링 문제에서 우수한 성능을 보이는 딥 러닝 모델
- 문장의 양방향 어순을 모두 사용
- 문장의 언어적 특성을 고려한 모델
- 출력 태그간 의존성을 고려하기 위해 output layer에 CRF 적용



제안 모델

Bi-LSTM-CRFs (의미역결정)

- 입력 어절에 대한 임베딩
  - 형태소 임베딩
  - 품사 태그 임베딩
  - 음절 포지션 인코딩
  - 음절 임베딩
  - **ELMo 임베딩**
    (학습 시, fine-tuning 되지 않음)
    - 첫번째 형태소의 ELMo 임베딩
    - 마지막 형태소의 ELMo 임베딩



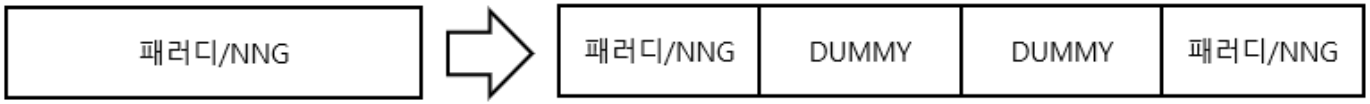
제안 모델

■ 어절 임베딩

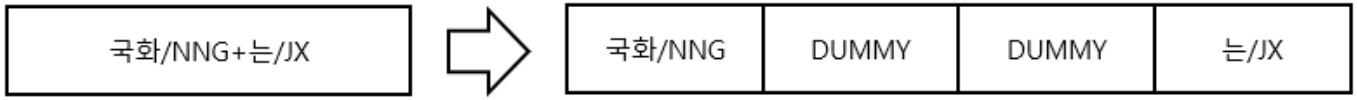
-> 형태소 임베딩의 조합으로 어절 표현

-> 4개의 형태소를 결합(concatenate)하여 사용

- 1) 한 개의 형태소로 구성된 어절



- 2) 두 개의 형태소로 구성된 어절



- 3) 세 개의 형태소로 구성된 어절

제안 모델

■ 어절 임베딩

-> 품사 태그 임베딩의 조합으로 어절 표현

-> 4개의 품사 태그를 결합(concatenate)하여 사용

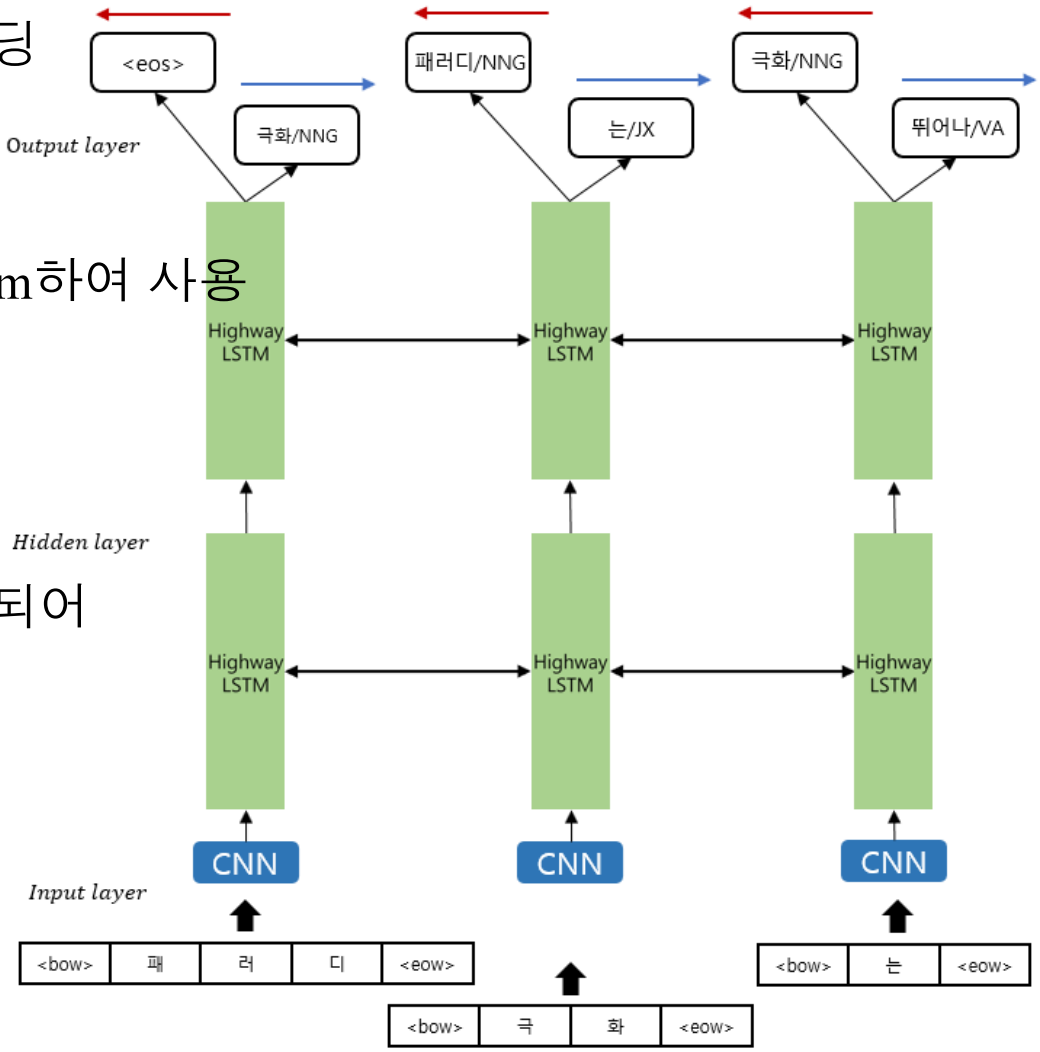
- 1) 한 개의 형태소로 구성된 어절

- 2) 두 개의 형태소로 구성된 어절

- 3) 세 개의 형태소로 구성된 어절

제안 모델

- **ELMo(Embedding From Language Model)**
  - 문맥 정보를 포함하고 있는 임베딩
  - Bi-LSTM Language Model
  - Highway LSTM 사용
  - LM의 Hidden state 를 weighted sum하여 사용

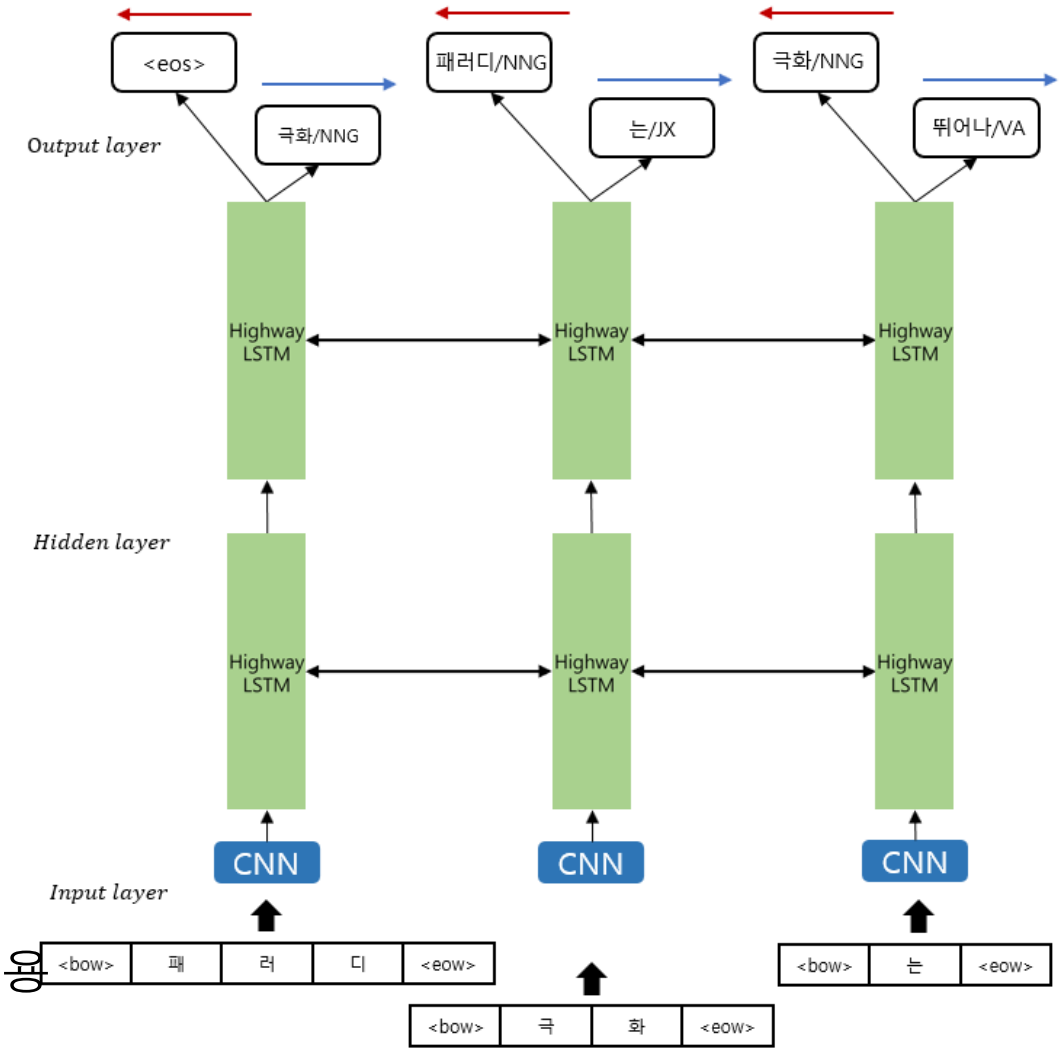


- **한국어 ELMo**
  - 형태소 단위 입출력 구조
  - 입력 형태소는 음절 단위로 분리되어 CNN을 거쳐 입력으로 사용됨
  - 다음 단어로 품사 태그가 포함된 형태소 예측

제안 모델

한국어 ELMo

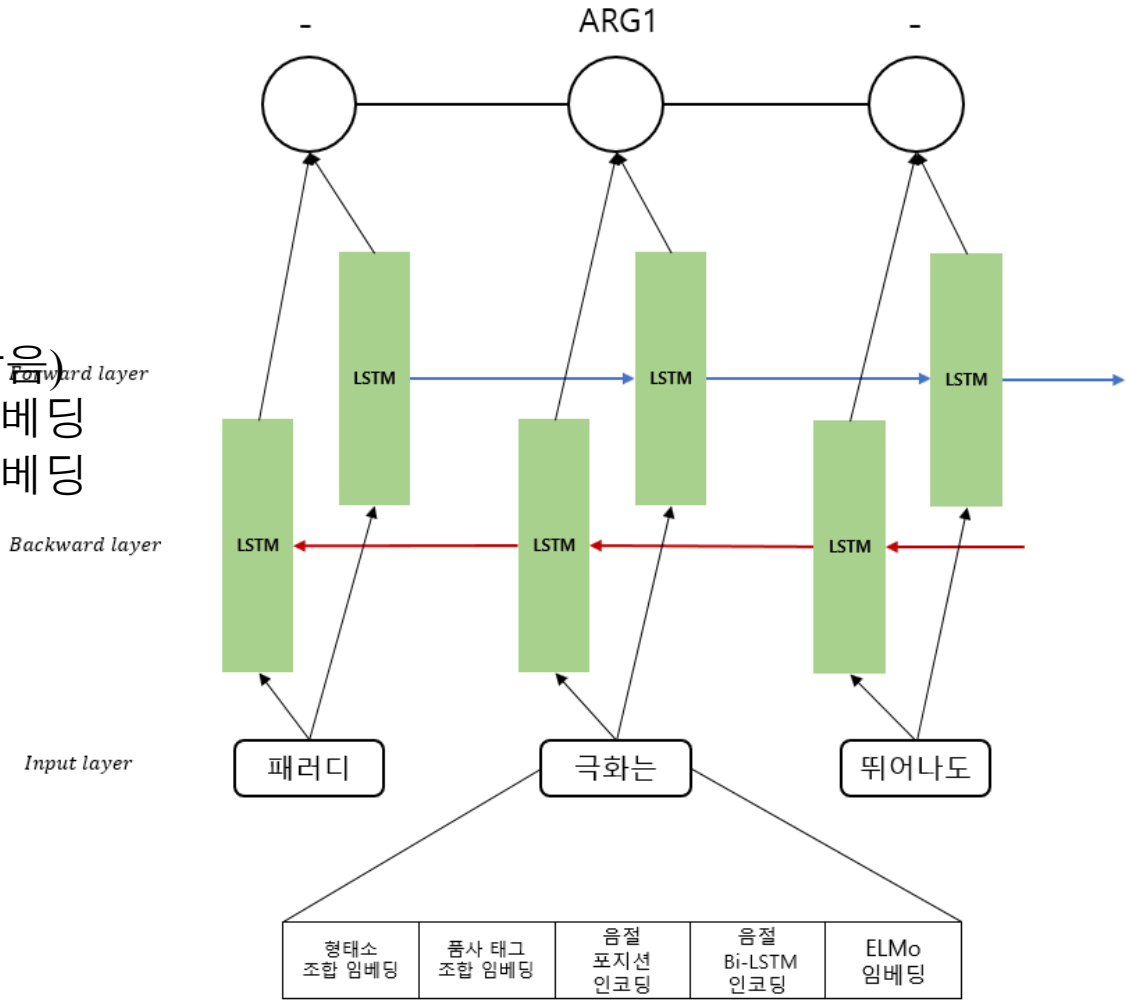
- 약 16GB 뉴스데이터를 형태소 분석하여 사용
  (Komoran 형태소 분석기)
  -> perplexity : 약 8.xx…
- 의미역결정 학습/검증 데이터를 사용하여 fine-tuning
  -> SRL perplexity : 약 1.xx…
  -> NER perplexity : 약 2.xx…
- 다음과 같은 symbol을 사용한 데이터 전처리 작업
  <bos> : begin of sentence.
  <eos> : end of sentence.
  <bow> : begin of word.
  <eow> : end of word.
- 1024차원의 ELMo embedding 사용



제안 모델

Bi-LSTM-CRFs (의미역결정)

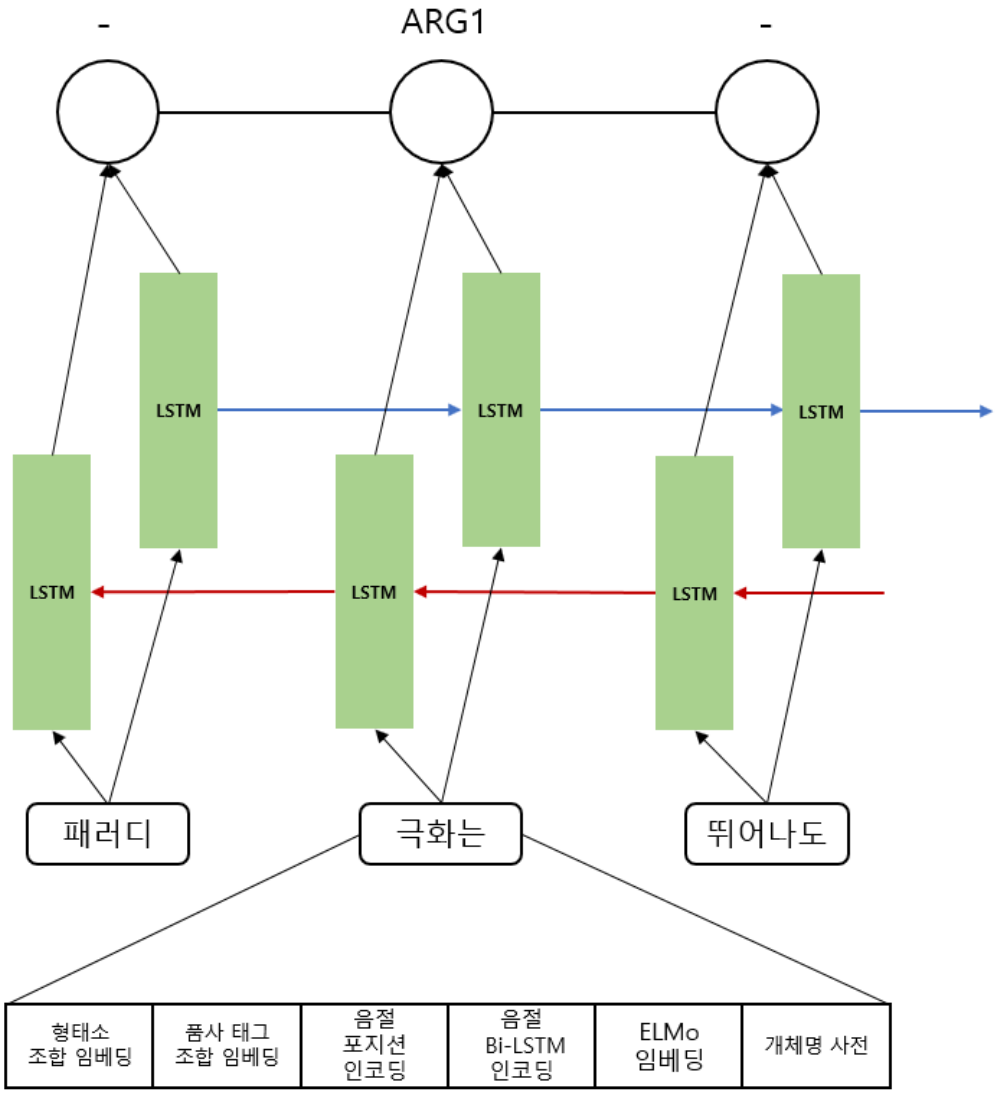
- 입력 어절에 대한 임베딩
  - 형태소 조합 임베딩
  - 품사 태그 조합 임베딩
  - 음절 포지션 인코딩
  - 음절 임베딩
  - **ELMo 임베딩**
    (학습 시, fine-tuning 되지 않음)
    - **첫번째** 형태소의 ELMo 임베딩
    - **마지막** 형태소의 ELMo 임베딩



제안 모델

Bi-LSTM-CRFs (개체명인식)

- 입력 어절에 대한 임베딩
  - 형태소 임베딩
  - 품사 태그 임베딩
  - 음절 포지션 인코딩
  - 음절 임베딩
  - ELMo 임베딩
    (학습 시, fine-tuning 되지 않음)
    - 첫번째 형태소의 ELMo 임베딩
    - 마지막 형태소의 ELMo 임베딩
  - **개체명 사전 자질**
    (챌린지 배포 사전)



실험

실험 데이터

- Komoran 형태소 분석기 사용

- 의미역 결정
  - 학습 데이터 : 31,856 문장
  - 검증 데이터 : 3,000 문장

- 개체명인식
  - 학습 데이터 : 81,000 문장
  - 검증 데이터 : 9,000 문장

실험

하이퍼파라미터

단어 임베딩	100 차원 (Xavier init)
음절 임베딩 품사 임베딩	50 차원 (Xavier init)
ELMo 사이즈	1024차원
LSTM 사이즈	200차원
LSTM 레이어	1
드랍아웃	0.7
배치사이즈	32
최적화 알고리즘	Adam
Learning rate	0.001

실험

성능 비교

- 의미역 결정

	Dev F1	Test F1
Bi-LSTM-CRFs	77.3	75.9
BI-LSTM-CRFs + ELMo	**78.1**	**77.6**

실험

■ 의미역 결정(SRL)

SRL

수상구분	Rank	Date	팀명	구성원	F1
대상	1	2018.12.14 22:01	Sogang_Alzzam	박찬민, 박영준 (서강대학교 자연어처리연구실)	77.6628
우수	2	2018.12.14 17:44	KANE_team	함영균, 김동환, 최기선 (KAIST SWRC)	76.3328
우수	3	2018.12.10 19:53	cheap_learning	박광현, 이영훈 (전북대학교)	76.2543
장려	4	2018.12.08 22:06	OnlyOne	김영천	75.4308
장려	5	2018.12.14 17:43	nlp_pln	이신의, 박장원, 박종성 (연세대학교 데이터공학연구실)	74.8749
장려	6	2018.12.14 09:17	kozistr_team	김형찬 (한국기술교육대학교)	74.7695

실험

■ 개체명인식(NER)

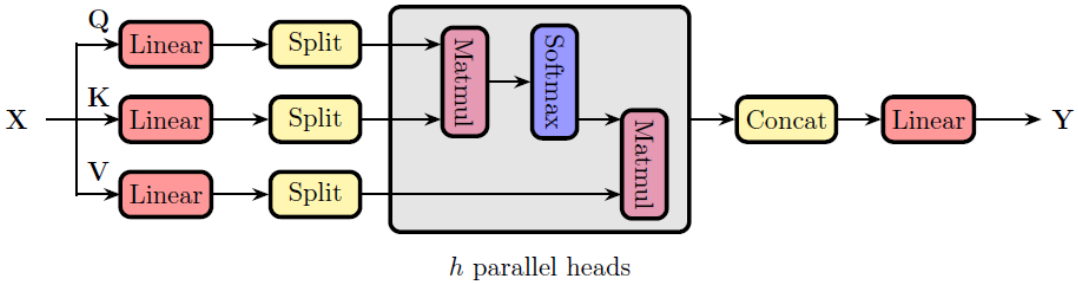
NER

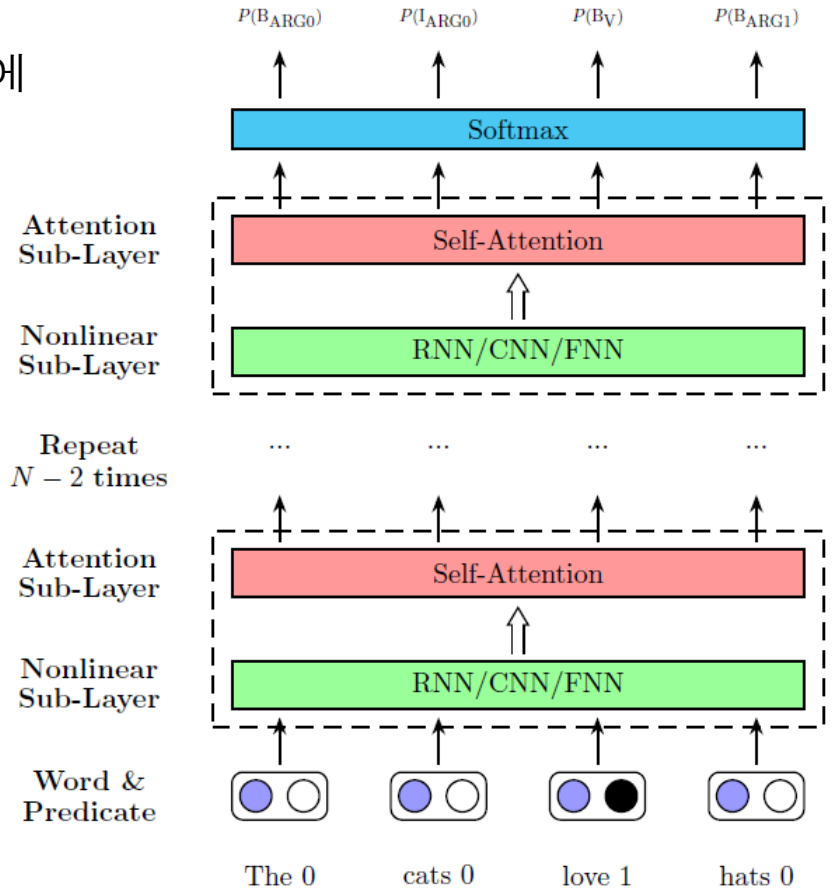
수상구분	Rank	Date	팀명	구성원	F1
대상	1	2018.12.14 07:45	State_Of_The_Art	박동주 (광주과학기술원)	90.4219
우수	2	2018.12.14 00:29	cheap_learning	박광현, 이영훈 (전북대학교)	90.2417
우수	3	2018.12.14 22:46	nlp_pln	이신의, 박장원, 박종성 (연세대학교 데이터공학연구실)	89.7830
장려	4	2018.12.14 15:18	Sogang_Alzzam	박찬민, 박영준 (서강대학교 자연어처리연구실)	88.8506
장려	5	2018.12.14 23:51	ner_master	조민수, 박찬희, 박진욱 (연세대학교 데이터공학연구실)	88.5818
장려	6	2018.12.13 19:28	bible	현청천 (HELLO NMS)	88.3348

실험

그 외 추가 실험

- **Self-attention**
  - Multi-head attention을 RNN의 output layer에 적용한 모델



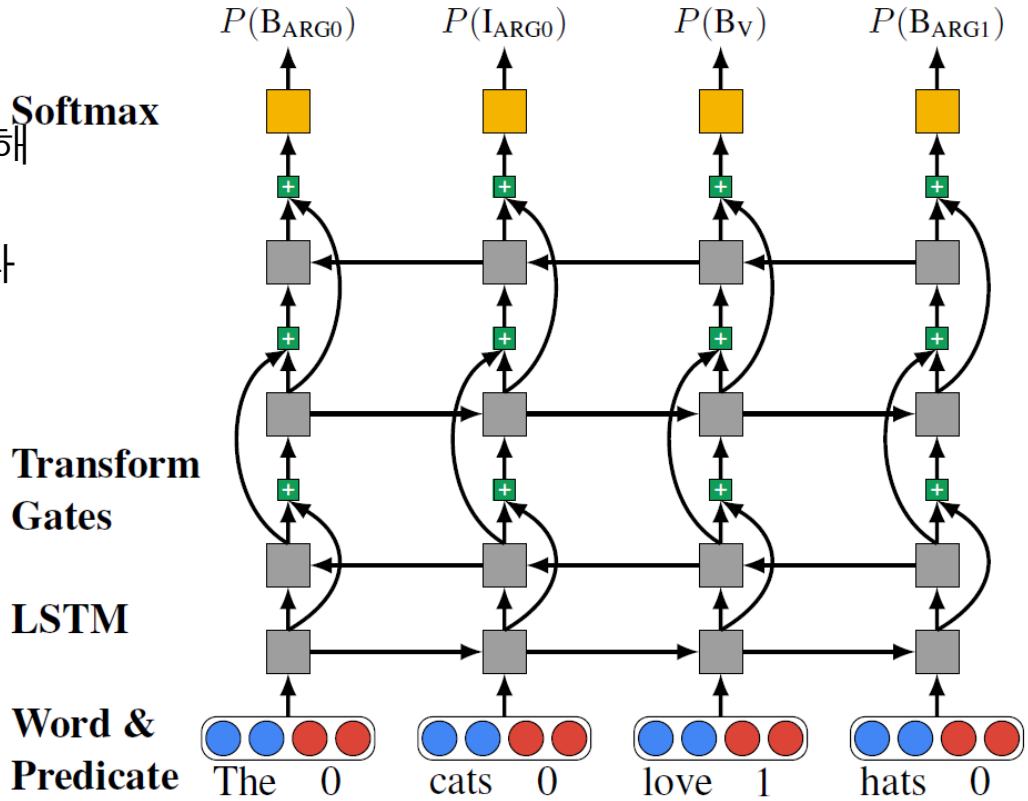


Tan, Zhixing, et al. "Deep semantic role labeling with self-attention." AAAI-2018.

실험

그 외 추가 실험

- **Highway-LSTM**
  - Residual connection의 일종
  - Vanishing gradient를 해결하기 위해 제안된 LSTM cell
  - 비선형변환을 거친 결과(output)와 거치지 않은 결과(raw input)를 gate 연산을 통해 계산

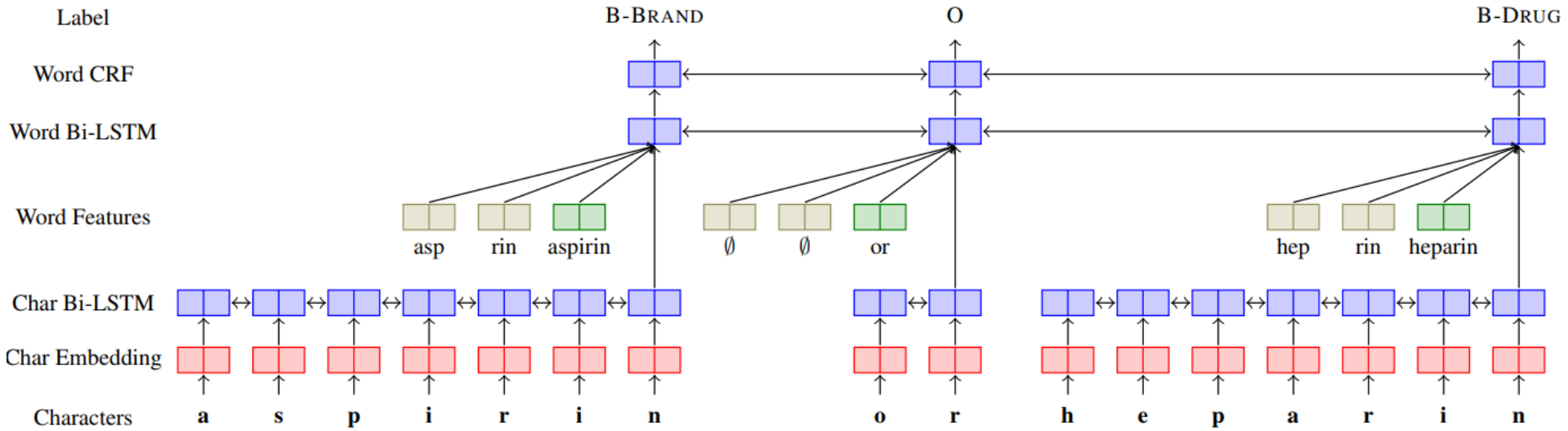


He, Luheng, et al. "Deep semantic role labeling: What works and what's next." *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 2017.

실험

그 외 추가 실험

- **Affix features**
  - 학습 데이터에서 빈도수가 높은 접두사/접미어를 추출하여 vocabulary 생성
  - 입력 형태소/어절에 대한 접두사/접미어를 추출하여 임베딩 학습



Yadav, Vikas, Rebecca Sharp, and Steven Bethard. "Deep Affix Features Improve Neural Named Entity Recognizers." *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. 2018.

결론

결론

- 한국어 특성에 적합한 ELMo 모델 학습
- 사전 학습된 ELMo를 의미역 결정, 개체명인식에 적용하여 성능 향상
- 향후 계획으로 한국어 ELMo 학습 시, 품사 태그 정보를 활용한 모델을 실험 예정

QnA

새해 복 많이 받으세요.

감사합니다.