심층 신경망(Deep Neural Network)을 이용한 자연어처리 도구 개발

2017.06.02

동아대학교 컴퓨터공학과

고영중

동아대학교 DONG-A UNIVERSITY

차례

1. 단추(Danchu): 언어분석 통합 툴
   - 단추 소개
   - 시스템 구성도

2. Bidirectional LSTM CRF

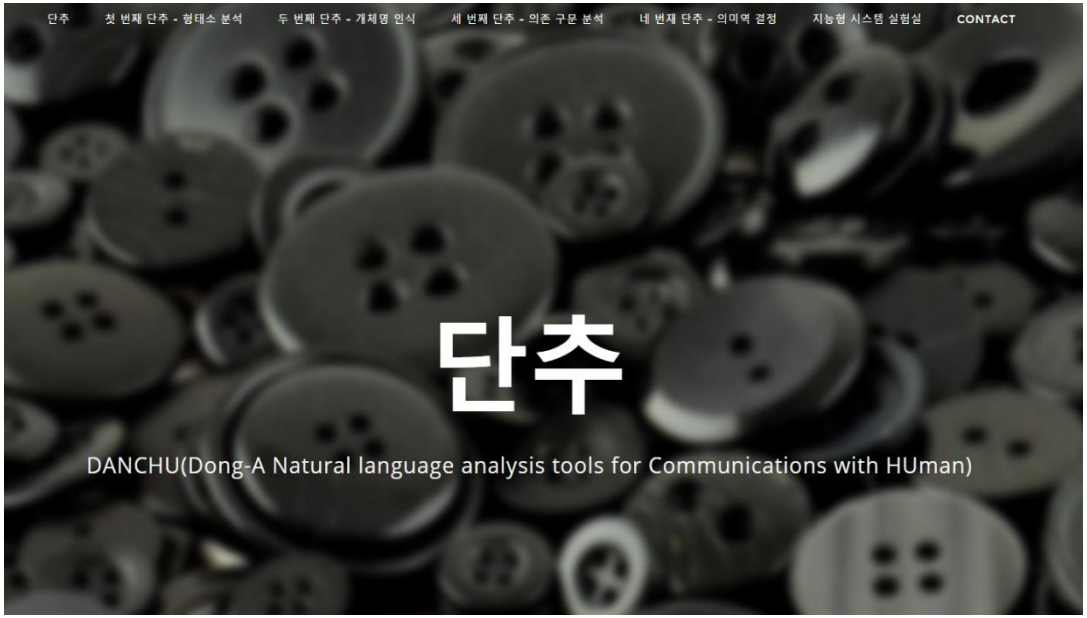
3. 언어 분석 기술 소개 및 성능 평가
   - 형태소 분석기
   - 개체명 인식기
   - 의존 구문 분석기
   - 의미역 결정기

4. 기계학습용 텍스트 데이터 레이블 자동생성 및 검증도구 개발

5. 요약

단추(Danchu) 소개

- 언어 분석 툴(단추) 개요
  - 자연어 처리의 시작
    - 첫 번째 단추를 잘 끼워야 함
    - 자연어 처리에서의 첫 번째 단추는 언어 분석
  - 딥 러닝 기반 모델
    - Bidirectional LSTM CRF
  - 포함 언어 분석기
    - 형태소 분석기
    - 개체명 인식기
    - 의존 구문 분석기
    - 의미역 결정기



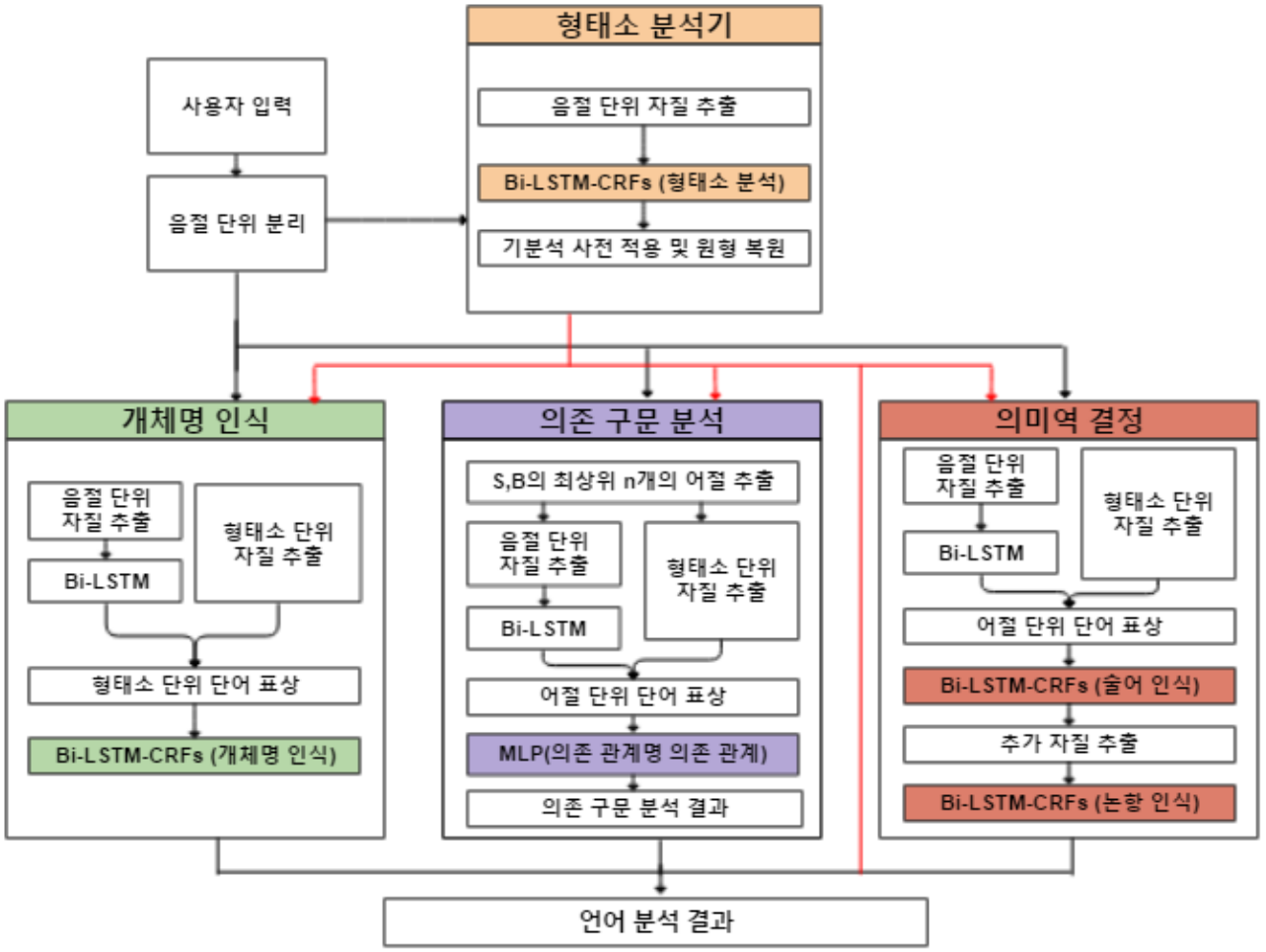
언어분석의 결과 예시

- 언어분석 결과 예시

형태소 분석	약/MM 3/SN 분간/NNG 의/JKG 예열/NNG 은/JX 엔진/NNG 을/JKO 보호하/VV ㅂ니다/EF ./SF
개체명 인식	약 3분/DT(시간)간의 예열은 엔진을 보호합니다. (애플/OG(기관)은 스티븐 잡스/PS(사람)에 의해 1976년/DT(시간) 창립 되었으며, 현재 본사는 미국 캘리포니아 주/LC(지역)에 있다.)
의존 구문분석	NP_SBJ (체언구_주어) DP (관형사구) / NP_MOD (체언구 체언 수식어) / NP_OBJ (체언구 목적어) 약 3분간의 예열은 엔진을 보호합니다.
의미역 결정	ARG0 (주동격) ARG1 (피동격) A1(동사) 약 3분간의 예열은 엔진을 보호합니다.

시스템 구성도

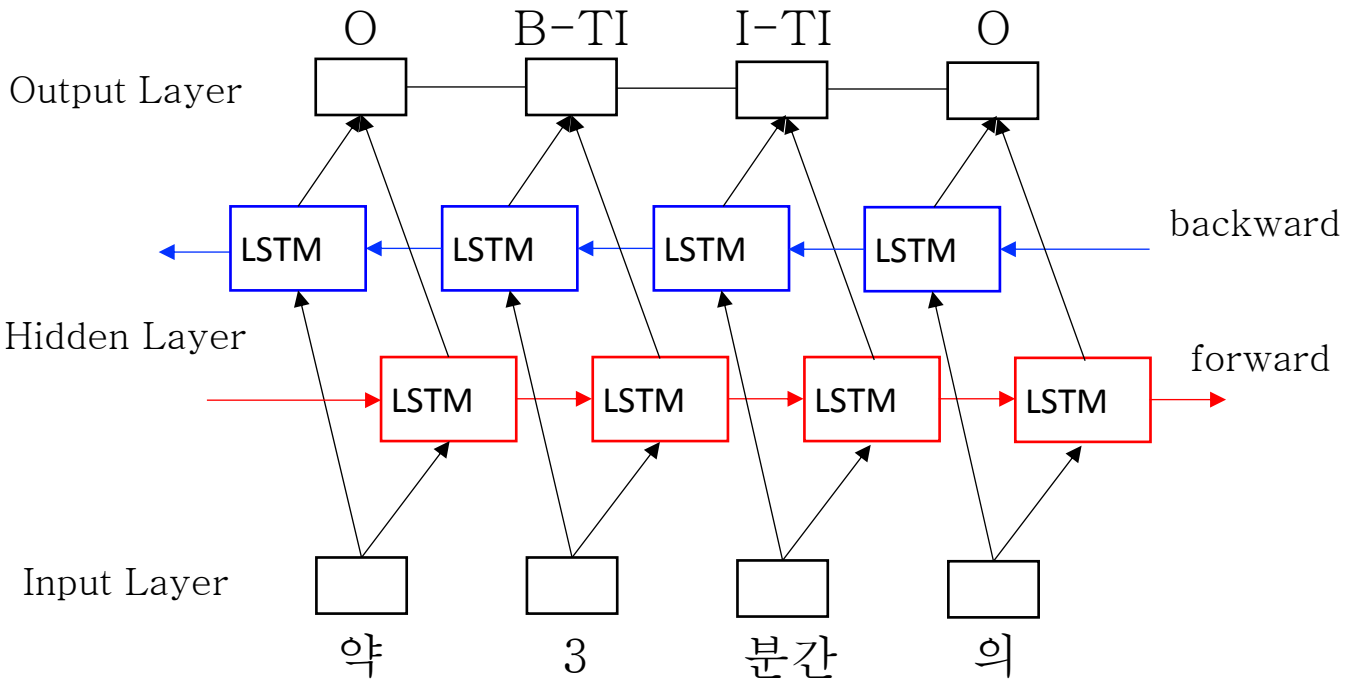
- 시스템 구성도



[시스템 전체 구성도]

Bidirectional LSTM CRF

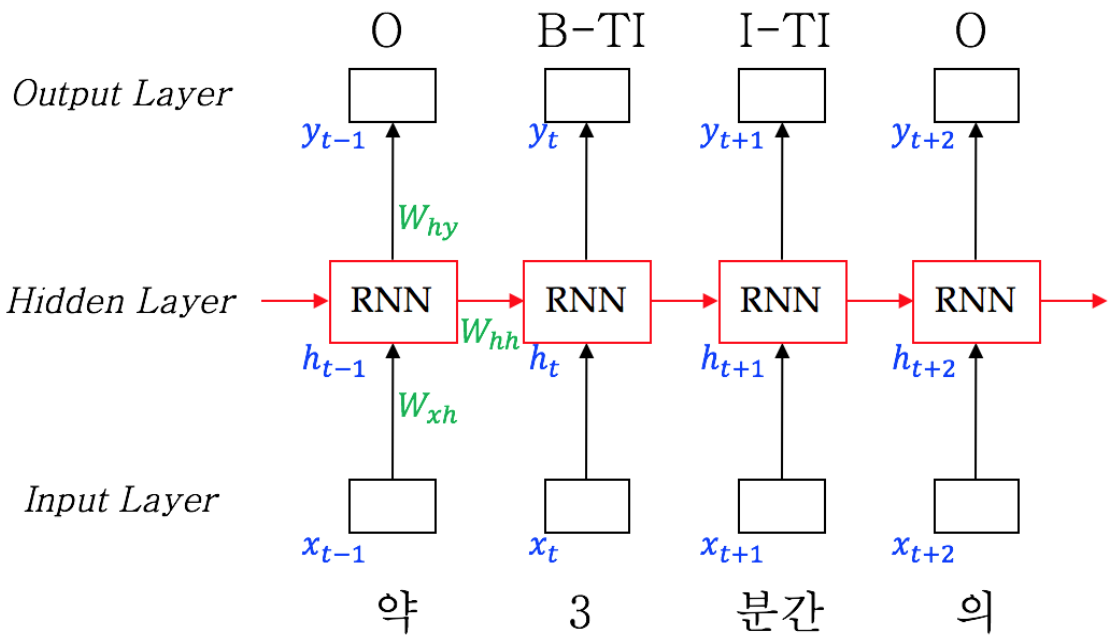
- **Bidirectional LSTM CRF**
  - **순차 레이블링에 특화된 LSTM기반 모델**
  - **언어 분석기는 순차 레이블링 시스템**



[Bidirectional LSTM CRF]

Bidirectional LSTM CRF

- **Recurrent Neural Networks(RNN)**
  - **LSTM의 기반이 되는 모델**
  - **장기 의존성을 유지하기 어려움**



[Recurrent Neural Networks]

$$h_t = \tanh(W_{xh}x_t + W_{hh}h_{t-1})$$

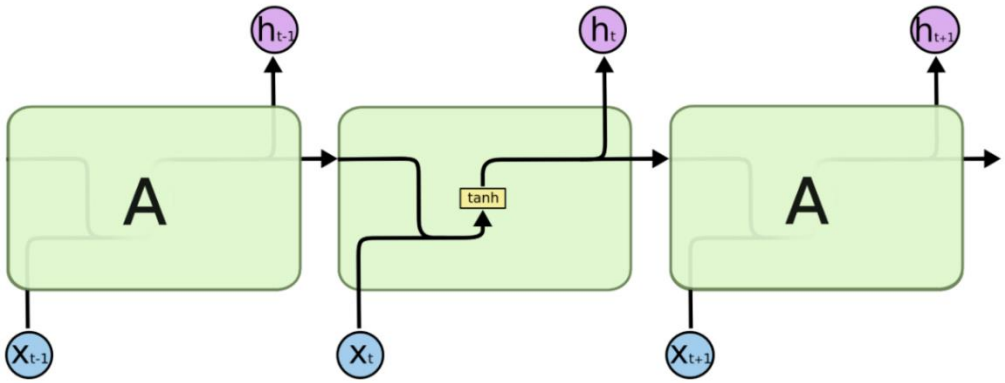
$$y_t = softmax(W_{hy}h_t)$$

Bidirectional LSTM CRF

- Long Shor-term Memory(LSTM)
  - RNN의 Vanishing Gradient Problem 해결
  - Memory cell
    - 전체적인 상태를 기억
  - Forget gate
    - Memory cell에서 어떤 정보를 제거할지 결정
  - Input gate
    - Memory cell에서 어떤 정보를 갱신할지 결정
  - Output gate
    - Memory cell에서 어떤 정보를 전달할지 결정

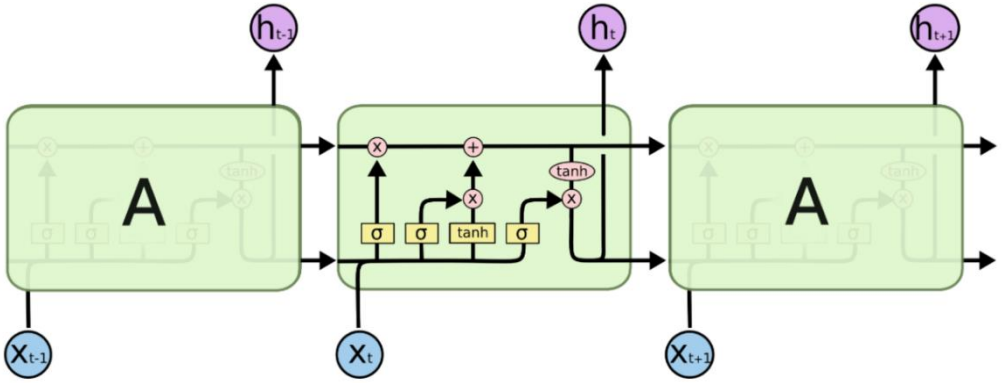
Bidirectional LSTM CRF

- **Long Shor-Term Memory(LSTM)**



[Recurrent Neural Networks]

$$h_t = \tanh(W_{xh}x_t + W_{hh}h_{t-1})$$



[Long Short-term Memory]

$$f_t = sigmoid\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right)$$

$$i_t = sigmoid(W_{xi}x_t + W_{hi}h_{t-1} + b_i)$$

$$\tilde{c}_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$

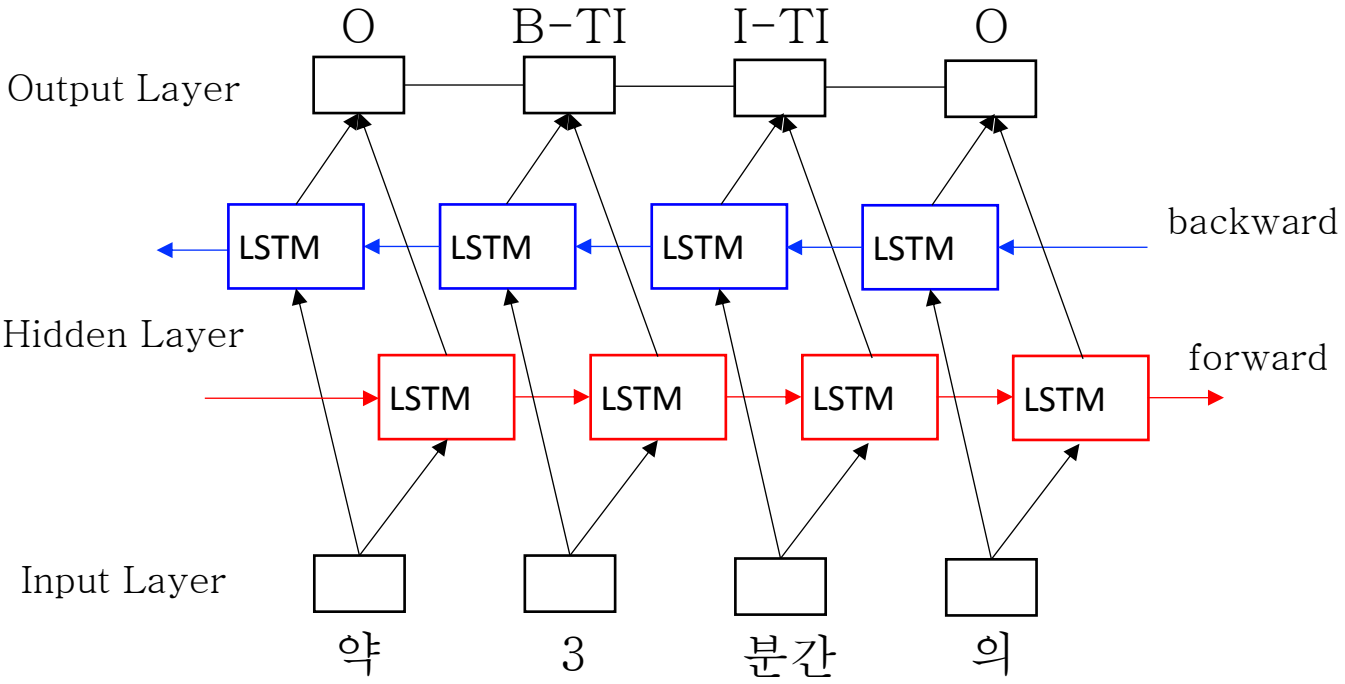
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

$$o_t = sigmoid(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$$

$$h_t = o_t \odot \tanh(c_t)$$

Bidirectional LSTM CRF

- **Bidirectional LSTM CRF**
  - **Output Layer의 정답 레이블간 의존성을 추가**
  - **전이 확률을 계산하기 위해 CRF의 forward 알고리즘 이용**
  - **최적의 열을 찾기 위해 Viterbi 알고리즘 이용**



[Bidirectional LSTM CRF]

언어 분석 기술

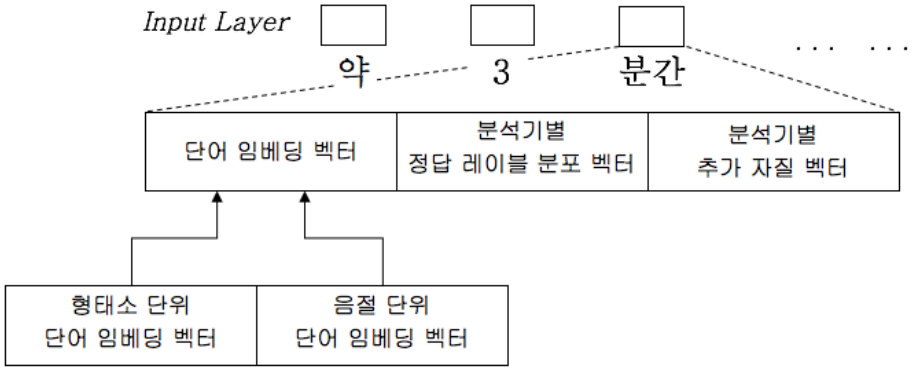
- 언어 분석기
  - 형태소 분석기
  - 개체명 인식기
  - 의존 구문 분석기
  - 의미역 결정기

  - 딥 러닝 기반 모델
    - Bidirectional LSTM CRF

  - 주요 입력 단어 표상
    - 단어 임베딩 벡터
    - 각 분석기별 정답 레이블 분포 벡터
    - 각 분석기별 추가 자질 벡터

입력 단어 표상(Word Representation)

- 주요 입력 단어 표상(Word Representation)
  - 단어 입베딩 벡터
    - 형태소 단위 임베딩 벡터
    - 음절 단위 임베딩 벡터
  - **분석기별 정답 레이블 분포 벡터**
    - 정답 레이블의 분포를 벡터로 표현
      - 각 분석기의 학습 데이터에서 추출
      - 형태소 및 음절 단위
  - 분석기 별 추가 자질 벡터
    - 각 분석기에 맞는 별도 자질 (형태소 태그 자질 등)



[이질적인 단어표상의 연결(Concatenation)]

단어 임베딩(Word Embedding)

- 단어 임베딩 벡터
  - Skip-gram을 이용한 학습
  - 11.5GB 뉴스 데이터 이용
    - 전체 약 22억 4,400만 형태소
    - Vocabulary Size : 약 191만 형태소

  - 단위 별 예시

[단어 임베딩 벡터 단위별 예시]

	예시	차원
원본 문장	약 3분간의 예열은 엔진을 보호합니다.	-
형태소 단위	약/MM, 0/SN, 분간/NNG, 의/JKG, 예열/NNG, 은/JX, 엔진/NNG, 을/JKO, 보호하/VV, ㅂ니다/EF, ./SF	64
음절 단위	약, 0, 분, 간, 의, 예, 열, 은, 엔, 진, 을, 보, 호, 합, 니, 다, .	32

정답 레이블 분포 벡터

- 정답 레이블 분포 벡터
  - 형태소 분석의 경우
    - 총 46차원 (품사태그 46개)
    - 한 단어가 각 품사 태그를 정답으로 가질 확률을 벡터로 만들어 사용
      - 학습 데이터에서 구축

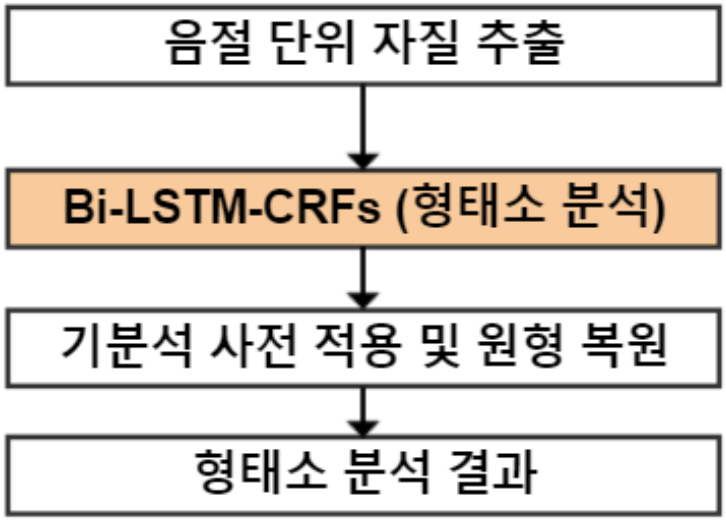
- 품사 분포 벡터 예시

[음절 단위 품사 분포 벡터 예시]

단위	단어	품사										
		NNG	NNP	NNB	VV	…	JKG	JKO	SN	…	MM	MAG
음절	약	0.49	0.01	0.00	0.03	…	0.00	0.00	0.00	…	**0.07**	0.00
	3	0.00	0.00	0.00	0.00	…	0.00	0.00	**0.91**	…	0.00	0.00
	분	**0.34**	0.01	0.08	0.07	…	0.00	0.00	0.00	…	0.00	0.00
	간	**0.46**	0.00	0.14	0.02	…	0.00	0.00	0.00	…	0.00	0.00
	의	0.08	0.00	0.00	0.02	…	**0.79**	0.00	0.00	…	0.00	0.00

형태소 분석기

- 형태소 분석
  - 음절 기반의 형태소 품사 태깅
  - 품사 : NNG, NNP, JKG, MM, SN, …



[형태소 분석 모델]

약, 3, 분, 간, 의, 예, 열, 은, 엔, 진, 을, 보, 호, 합, 니, 다, .

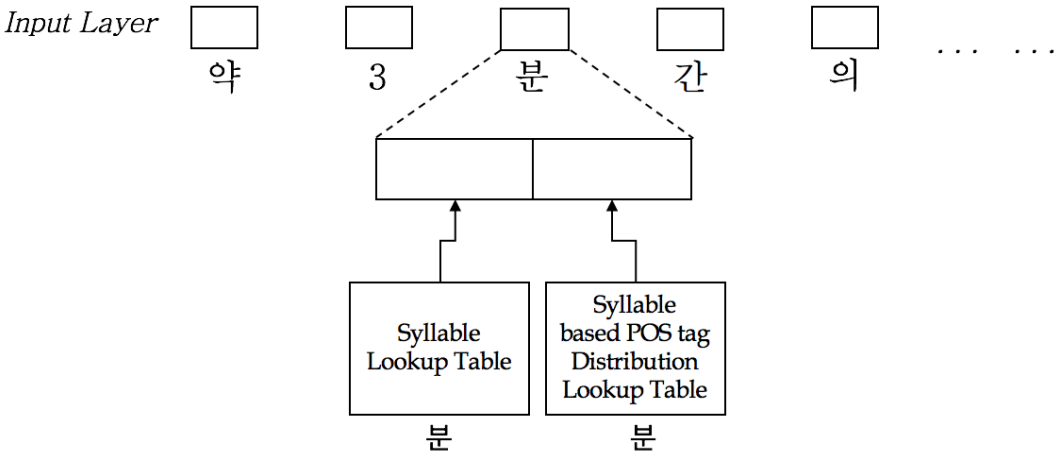
약/B-MM, 3/B-SN, 분/B-NNG, 간/I-NNG, 의/B-JKG, …

약/MM, 3/SN+분간/NNG+의/JKG, …

약/MM
3/SN+분간/NNG+의/JKG
예열/NNG+은/JX
엔진/NNG+을/JKO
보호하/VV+ㅂ니다/EF+./SF

형태소 분석기

- 형태소 분석
  - 입력 단어 표상(Word Representations)



[형태소 분석 입력 단어표상]

성능 평가

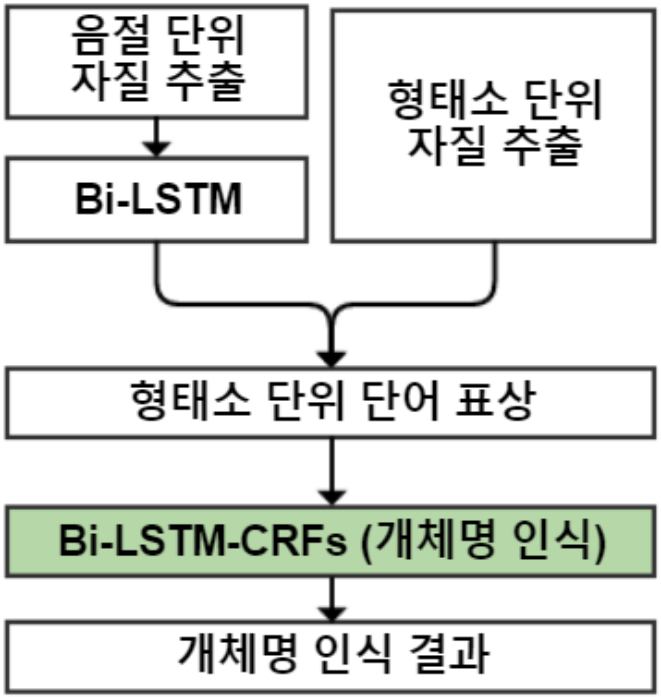
- 실험 환경
  - 세종 말뭉치
    - 8,376 문장 (학습 6,701 문장, 평가 1,675 문장)
    - 46 레이블

[형태소 분석 성능]

	Accuracy(%)	F1(%)
이건일 외 (2017)	95.40	96.91
나승훈 (2012)	-	96.19
심광섭 (2013)	96.60	-
황현선 외(2016)	-	97.08
이창기 (2013)	97.96	98.03
나승훈 외 (2014)	98.23	-
이충희 외 (2016)	99.03	-
단추 (Ours)	98.04	98.65

개체명 인식기

- 개체명 인식
  - 음절 및 형태소 기반의 개체명 인식
  - 개체명 : PS(인명), LC(지명), OG(조직명), DT(날짜), TI(시간)



[개체명 인식 모델]

약, 3, 분, 간, 의, …

약/MM 3/SN+분간/NNG+의/JKG, …

약, 3분간, 의, 예열, 은, 엔진, 을, 보호하, ㅂ니다, .

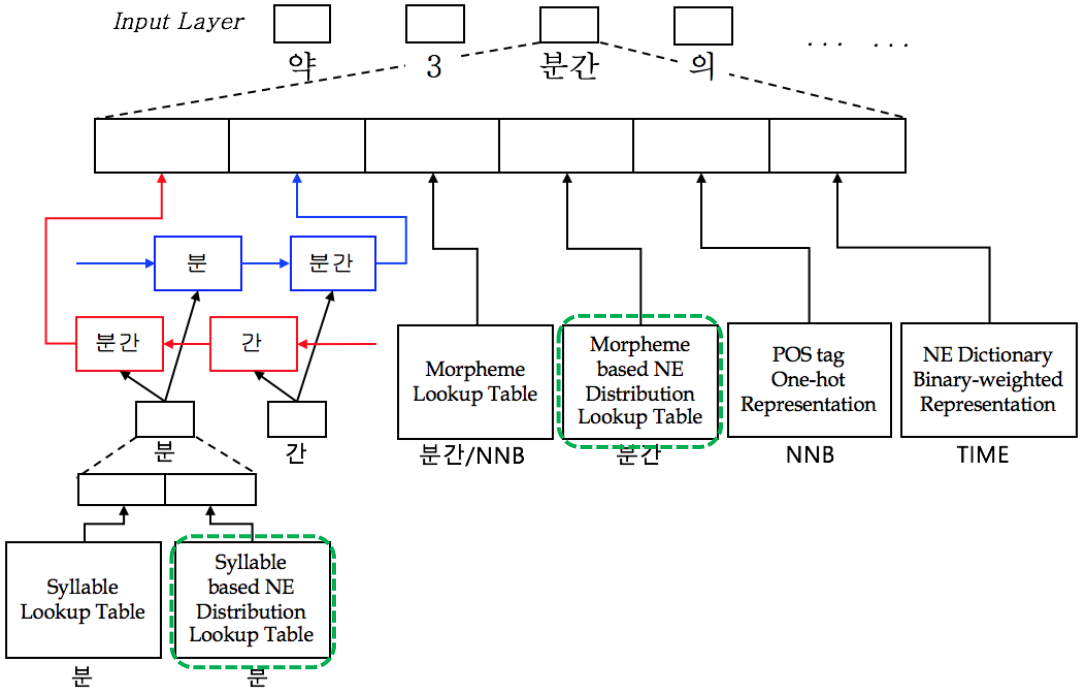
약/O 3/B-TI+분간/I-TI+의/O, …

TIME

약 3분간의 예열은 엔진을 보호합니다.

개체명 인식기

- 개체명 인식
  - 입력 단어 표상(Word Representations)
    - 음절 기반 형태소 단위 입력
      - Bidirectional LSTM을 이용한 확장
    - 형태소 단위 입력



[개체명 인식 입력 단어표상]

성능 평가

- 실험 환경
  - 2016 국어경진대회 말뭉치
    - 4,555 문장 (학습 3,555 문장, 평가 1,000 문장)
    - 5 레이블

[한국어 개체명 인식 성능]

	F1(%)		
	in-domain	out-domain	6:4
KAISER (2016)	58.94	45.49	53.56
Wordangler (2016)	78.07	62.58	71.87
Annie (2016)	84.17	63.01	75.70
서강 알짱 (2016)	87.62	70.79	80.89
KoNER (2016)	85.76	73.83	81.00
단추(Ours)	87.08	76.54	82.86

[영어 개체명 인식 성능]

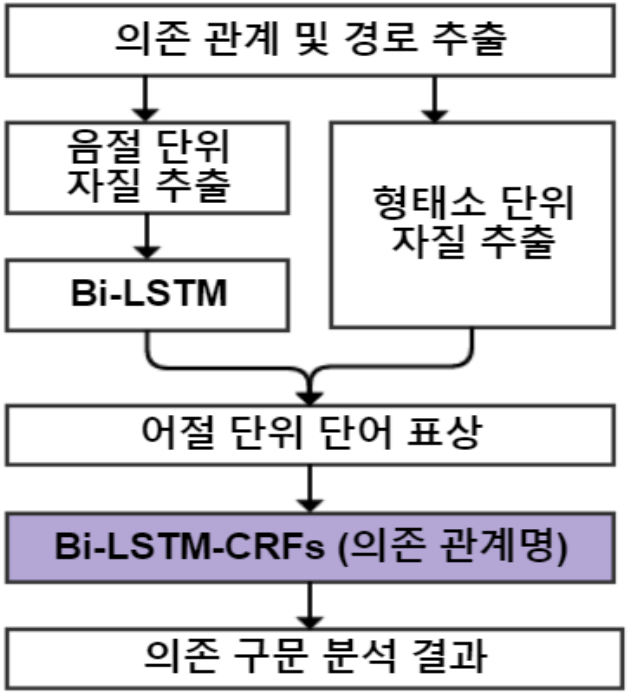
	F1(%)
Collobert et al. (2011)	89.59
Huang et al. (2015)	90.10
Chiu and Nichols (2015)	90.77
Ratinov and Roth (2009)	90.80
Lin and Wu (2009)	90.90
Passos et al. (2014)	90.90
Lample et al. (2016)	90.94
Luo et al. (2015)	91.20
Ma and Hovy (2016)	91.21
단추(Ours)	91.37

의존 구문 분석기

- 의존 구문 분석
  - 기존 의존 구문 분석 모델
    - 어절간의 의존 관계 분석에 초점 (UAS(Unlabeled Attachment Score)로 평가)
  - 최근 의존 관계 뿐만 아닐 의존명까지 동시 분석
    - LAS(Labeled Attachment Score)로 평가
  - 의존 관계명 분석 모델, 의존 관계 및 의존 관계명 동시 분석 두 가지 모델 보유
    - 의존 경로를 이용한 의존 관계명 분석 모델 (Bidirectional LSTM CRF)
    - 의존 관계 및 의존 관계명 분석 모델 (Multi-layer Perceptron(MLP))
  - Transition-based 의존 구문 분석(MLP)
    - Arc–Eager Transition, Backward 기반 모델

의존 구문 분석기

- **음절 및 어절 기반의 의존 관계명 분석 모델**

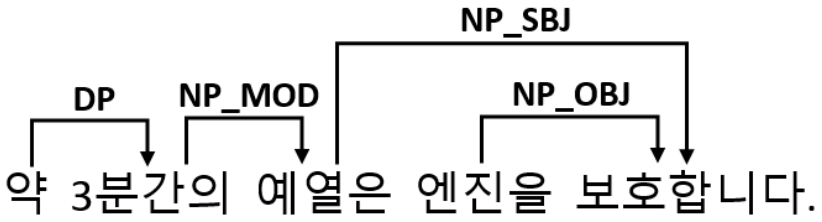


[의존 관계명 분석 모델]

약 → 3분간의 → 예열은 → 보호합니다.
엔진을 → 보호합니다.

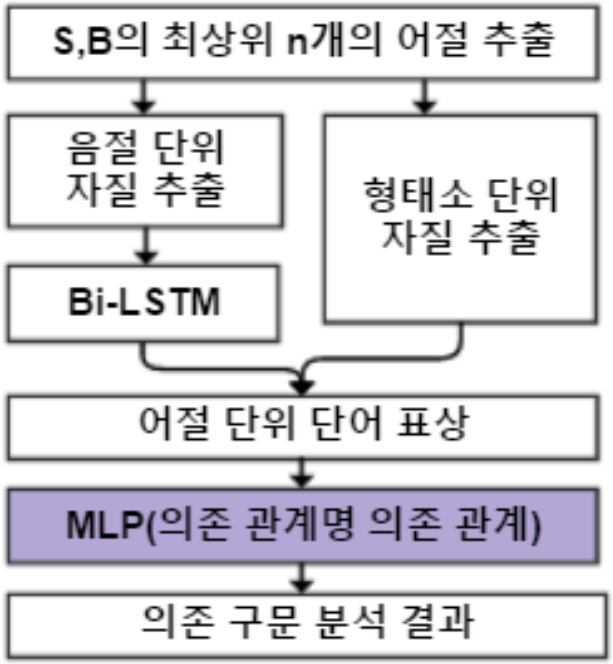
약, 3, 분, 간, 의, ...
약/MM 3/SN+분간/NNG+의/JKG, ...

약, 3분간의, 예열은, 엔진을, 보호합니다.

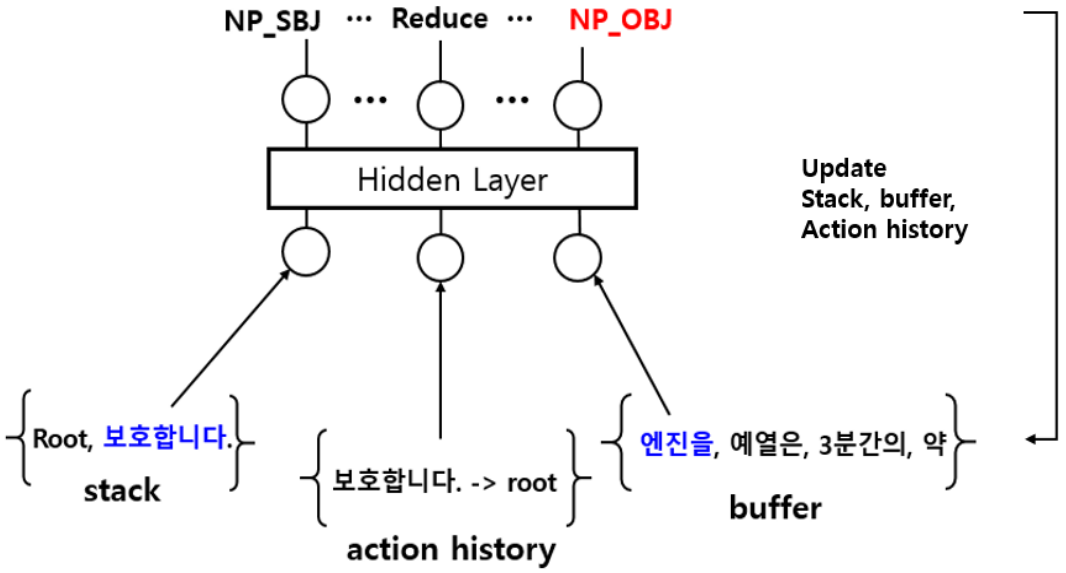


의존 구문 분석기

- 의존 관계 및 의존 관계명 분석 모델



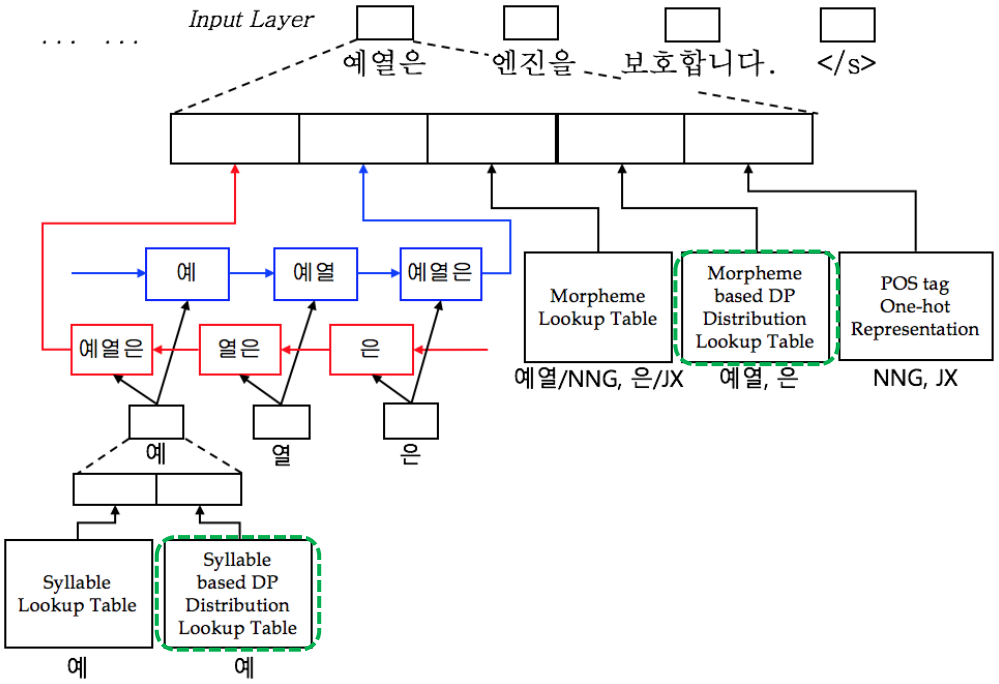
[의존관계 및 의존관계명 동시 분석 모델]



[MLP 기반 컨트롤 네트워크 구성도]

의존 구문 분석기

- 의존 구문 분석
  - 입력 단어 표상(Word Representations)
    - 음절 기반 어절 단위 입력
      - Bidirectional LSTM을 이용한 확장
    - 어절 단위 입력



[의존 구문 분석 입력 단어표상]

성능 평가

- 실험 환경
  - 세종 말뭉치
    - 59,574 문장 (학습 53,757 문장, 평가 5,817 문장)
    - 36 레이블

[의존 관계명 부착 성능]

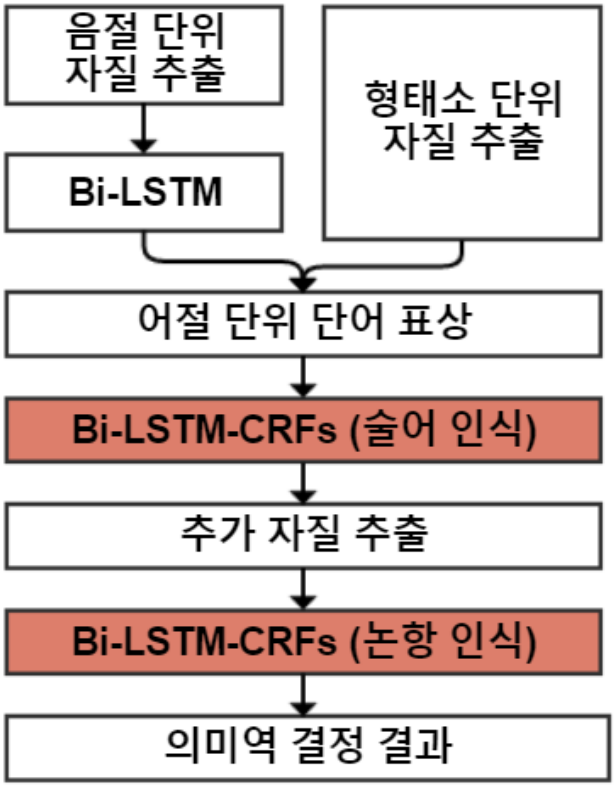
	F1(%)
정석원 외 (2013)	90.80
단추(Ours)	96.01

[의존 구문 분석 성능]

	UAS	LAS
오진영 외 (2013)	85.61	-
안광모 외 (2014)	87.52	-
이창기 외 (2014)	90.37	88.17
나승훈 외 (2016)	90.69	88.56
단추(Ours)	87.92	84.56

의미역 결정기

- 의미역 결정
  - 음절 및 어절 기반의 의미역 결정
  - 의미역 : ARG0, ARG1, ARGM-TMP, O, …



[의미역 결정 모델]

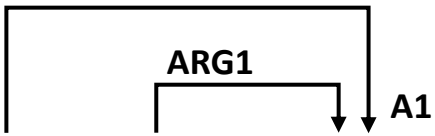
약, 3, 분, 간, 의, …

약/MM 3/SN+분간/NNG+의/JKG, …

약, 3분간의, 예열은, 엔진을, 보호합니다.

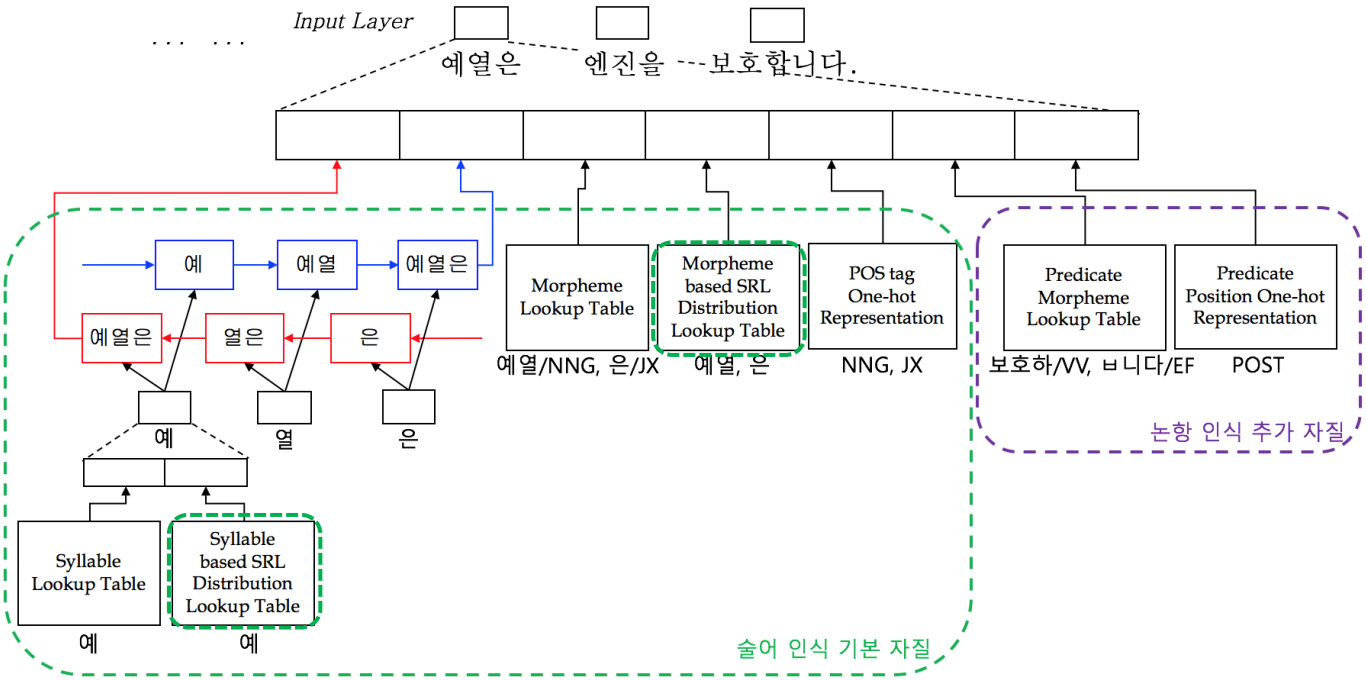
약/O, 3분간의/O, …, 보호합니다./A1

보호하/VV+ㅂ니다/EF



의미역 결정기

- 의미역 결정
  - 입력 단어 표상(Word Representations)
    - 음절 기반 어절 단위 입력
      - Bidirectional LSTM을 이용한 확장
    - 어절 단위 입력



[의미역 결정 입력 단어표상]

성능 평가

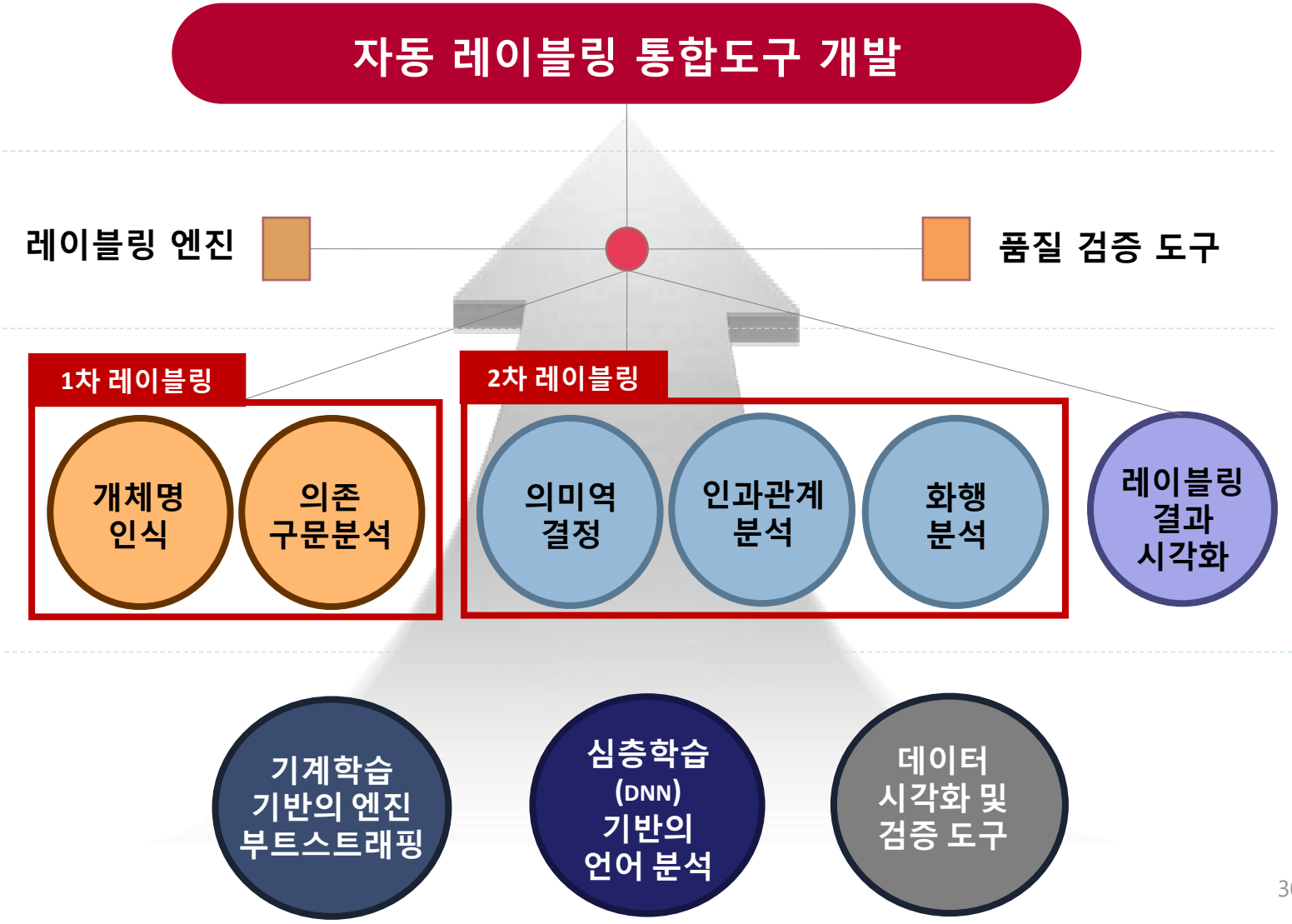
- 실험 환경
  - Korean PropBank
    - 4,853 문장 (학습 3,883 문장, 평가 970 문장)
    - 23 레이블

[의미역 결정 성능]

	F1(%)
이창기 외 (2015)	76.96
임수종 외 (2015)	79.54
배장성 외 (2015)	78.17
배장성 외 (2017)	78.57
단추(Ours)	80.24

기계학습용 텍스트 데이터 레이블 자동생성 및 검증도구 개발

연구 목표

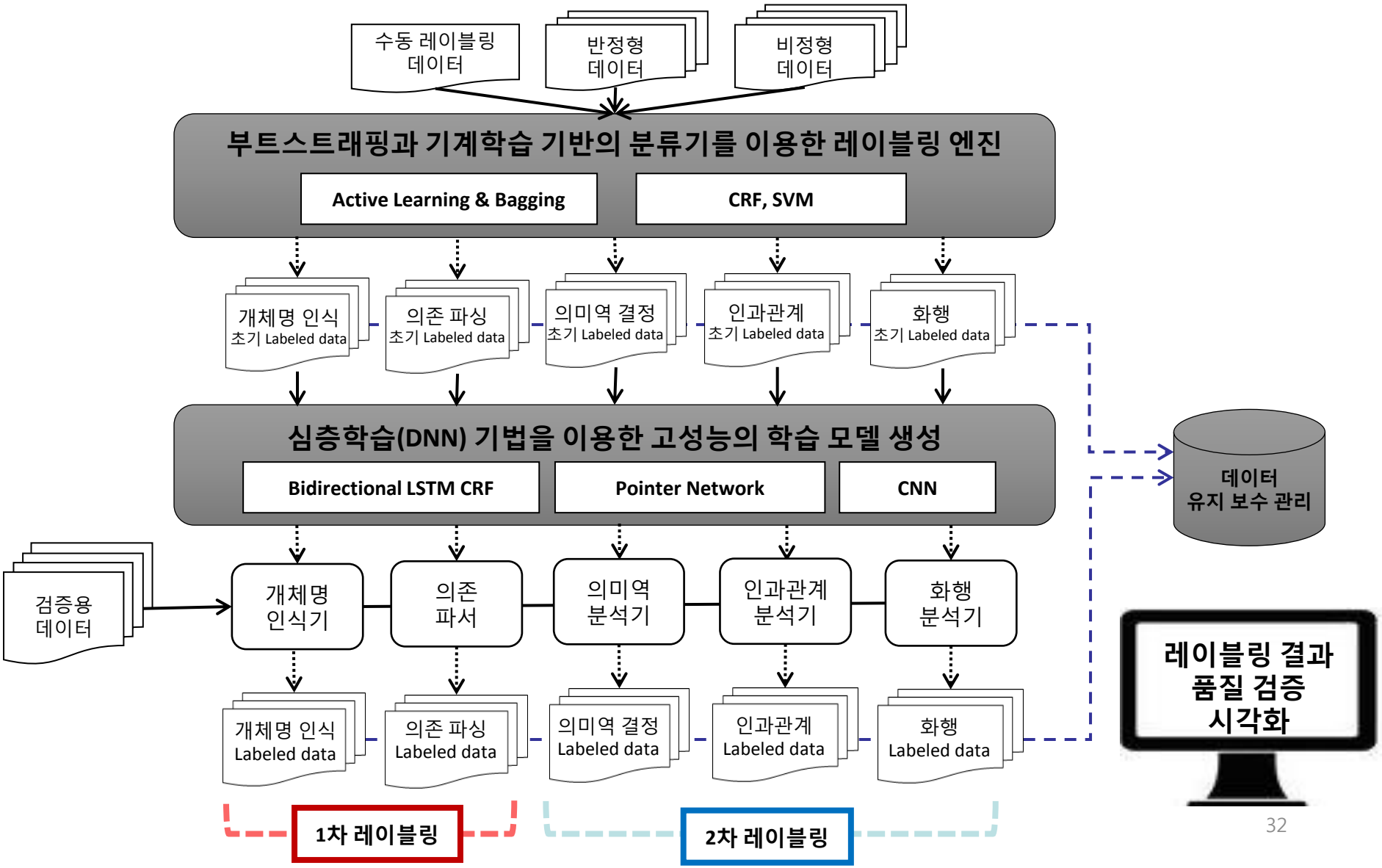

자동 레이블링 통합도구 개발
레이블링 엔진
품질 검증 도구
1차 레이블링
개체명 인식
의존 구문분석
2차 레이블링
의미역 결정
인과관계 분석
화행 분석
레이블링 결과 시각화
기계학습 기반의 엔진 부트스트래핑
심층학습 (DNN) 기반의 언어 분석
데이터 시각화 및 검증 도구

자동 레이블링 기반 기술 및 예시

- **레이블링 예시**

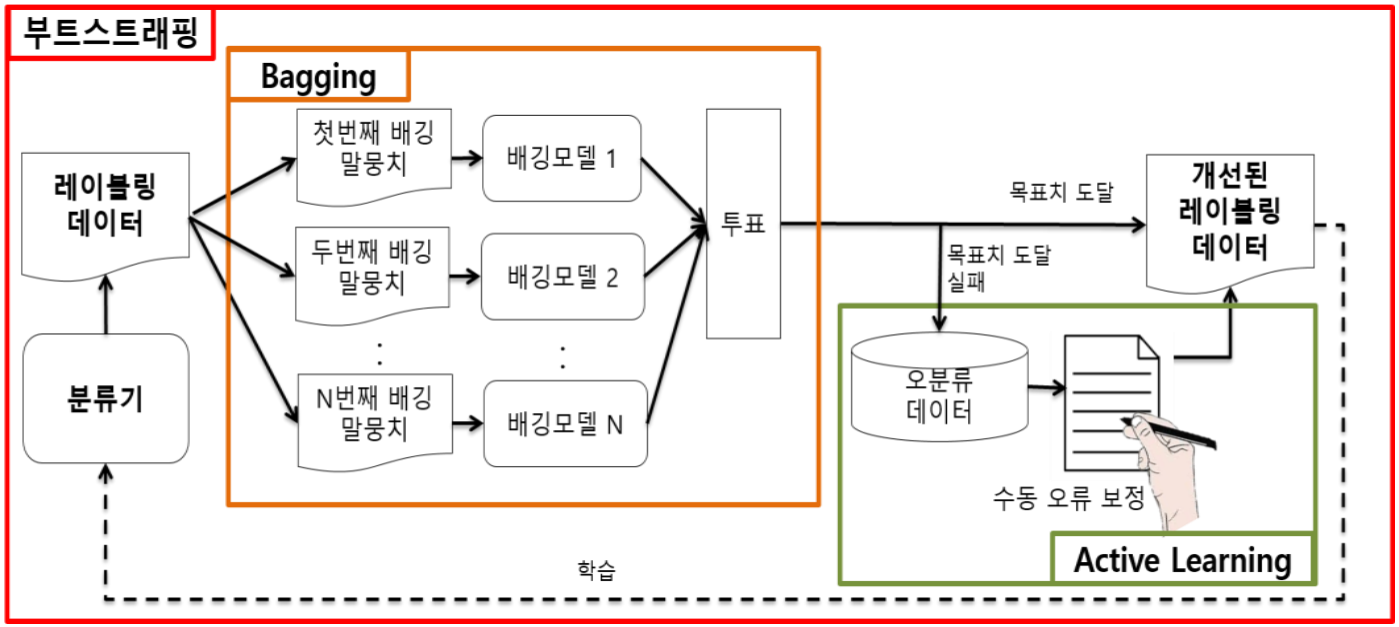
인과관계 분석	cause effect [약 3분간의 예열은] [엔진을 보호합니다.]
화행 분석	A : 안녕하세요? — Opening B : 네, 안녕하세요. 반가워요! — Opening A : 오늘 날씨가 참 좋아요. — Expressive B : 네. 하지만 오늘 오후에 비가 온대요. — Inform A : 그렇다면 우산을 챙겨야 겠군요! — Ask-confirm B : 네. 그러는 것이 좋을 것 같아요. — Accept

자동 레이블링 통합 도구 전체 구상도

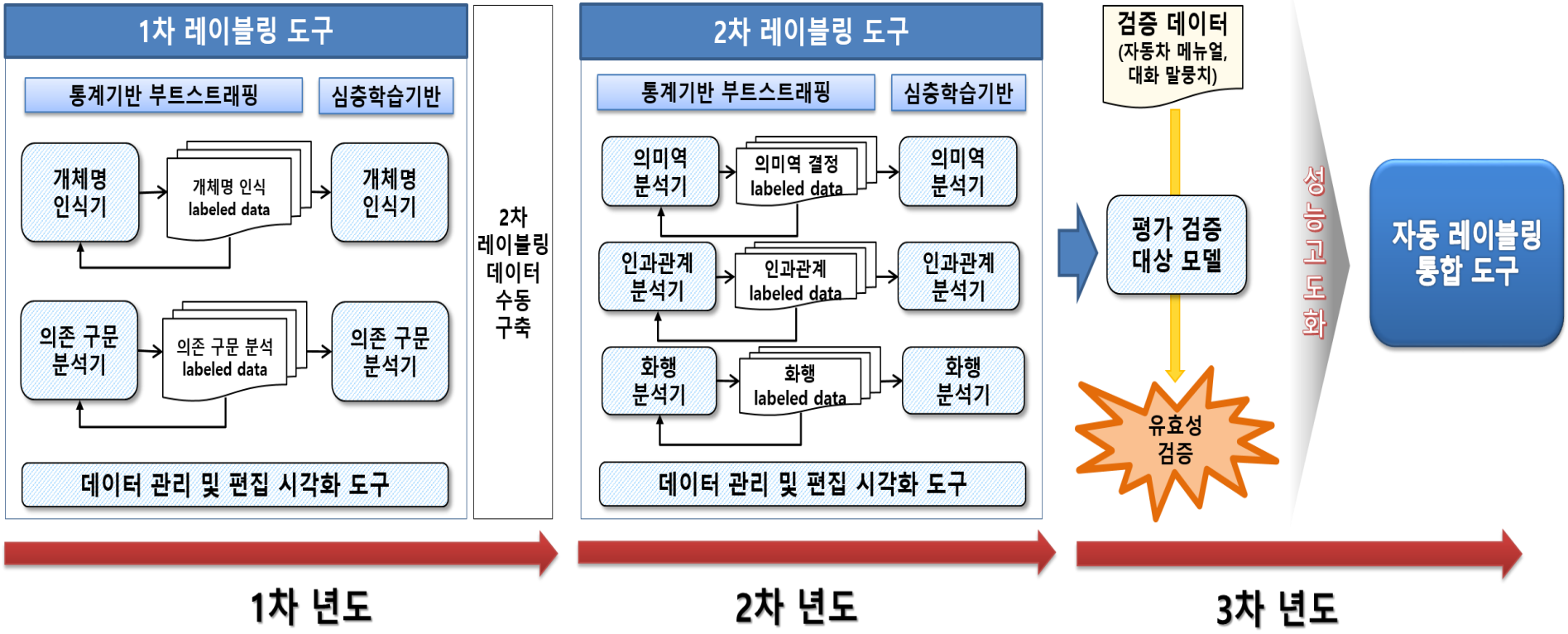


기계학습 기반 부트스트래핑을 활용한 자동 레이블링 기술

- 부트스트래핑 알고리즘 적용
  - 반복적 학습으로 분류기 성능 향상
  - 대량의 언레이블링 데이터로부터 레이블링 데이터 자동 생성
- 레이블링 오류 자동 교정
  - 오류의 학습을 방지
  - Bagging : 다수의 학습기로 부터 투표를 통해 오류 감소
  - Active Learning : 수동 교정으로 오류 감소



연차별 연구 내용


1차 레이블링 도구
통계기반 부트스트래핑
심층학습기반
개체명 인식기
개체명 인식 labeled data
개체명 인식기
의존 구문 분석기
의존 구문 분석 labeled data
의존 구문 분석기
데이터 관리 및 편집 시각화 도구
2차 레이블링 데이터 수동 구축
2차 레이블링 도구
통계기반 부트스트래핑
심층학습기반
의미역 분석기
의미역 결정 labeled data
의미역 분석기
인과관계 분석기
인과관계 labeled data
인과관계 분석기
화행 분석기
화행 labeled data
화행 분석기
데이터 관리 및 편집 시각화 도구
검증 데이터 (자동차 메뉴얼, 대화 말뭉치)
평가 검증 대상 모델
유효성 검증
성능 고도화
자동 레이블링 통합 도구
1차 년도
2차 년도
3차 년도

연구 추진 체계


전체 연구 조정 및 협업 주도
(서강대학교, 동아대학교, 다이퀘스트)
<기계학습 기반 기술>
부트스트래핑
서강대학교
동아대학교
기계학습 기반 모델
서강대학교
기계학습 모델 성능 고도화
서강대학교
<심층 학습 기술>
심층 학습 기반 모델
서강대학교
동아대학교
심층 학습 모델 성능 고도화
서강대학교
동아대학교
<데이터 관리 및 시각화>
데이터 관리 도구
다이퀘스트
편집 시각화 도구
다이퀘스트
<연구 결과물 활용>
공개 SW 운영
데이터 공유
서강대학교

파급 효과

반정형/비정형 데이터

자동 레이블링 통합 도구

자동 레이블링 엔진

레이블링 데이터

데이터 통합 관리 및 시각화

대화 시스템

빅데이터/자연어처리 연구

질의응답 시스템

인공지능 발전

요 약

- 같은 심층학습 모델을 형태소 분석, 의존 분석, 개체명 인식, 의미역 결정기에 적용하여 모든 분석기에서 높은 성능 획득
  - 워드/음절 임베딩 + 정답 레이블 분포 벡터
  - Bidirectional LSTM CRF

- 텍스트 데이터 레이블 자동생성 및 검증도구 개발
  - Booststrapping (Bagging, Active Learning)
  - Bidirectional LSTM CRF, Pointer Network, Convolutional Neural Network
  - 향후 툴을 오픈해서 인공지능 발전에 기여

- 향후 연구
  - 심층학습 기반 대화 시스템 개발
    - SLU, State Tracker, Dialogue Policy Manager
  - 심층학습 기반 인과관계 분석기

Q&A