단추(DANCHU)

유홍연, 안재현, 윤정민, 김혜민, 박용신, 배경만, 고영중

지능형시스템 연구실

동아대학교

목차

- 인터페이스
- 개요 및 개발 동기
- 시스템 아키텍처
- 목표 및 특징
- 프로그램 기능
- 제안 방법
- 성능 및 평가
- 참고 문헌
- 결론

개요 및 개발 동기

DANCHU 개요

- *Dong-A Natural language analysis tools for Communication with HUman*
- 언어분석 통합 시스템

DANCHU 개발 동기

- 인간이 발화하는 언어현상을 기계적으로 분석하여 기계가 이해할 수 있는 형태로 만드는 것을 의미
- “ 잘 분석된 언어 분석은 많은 자연어 응용 분야에서 좋은 토대”

목표 및 특징

목표

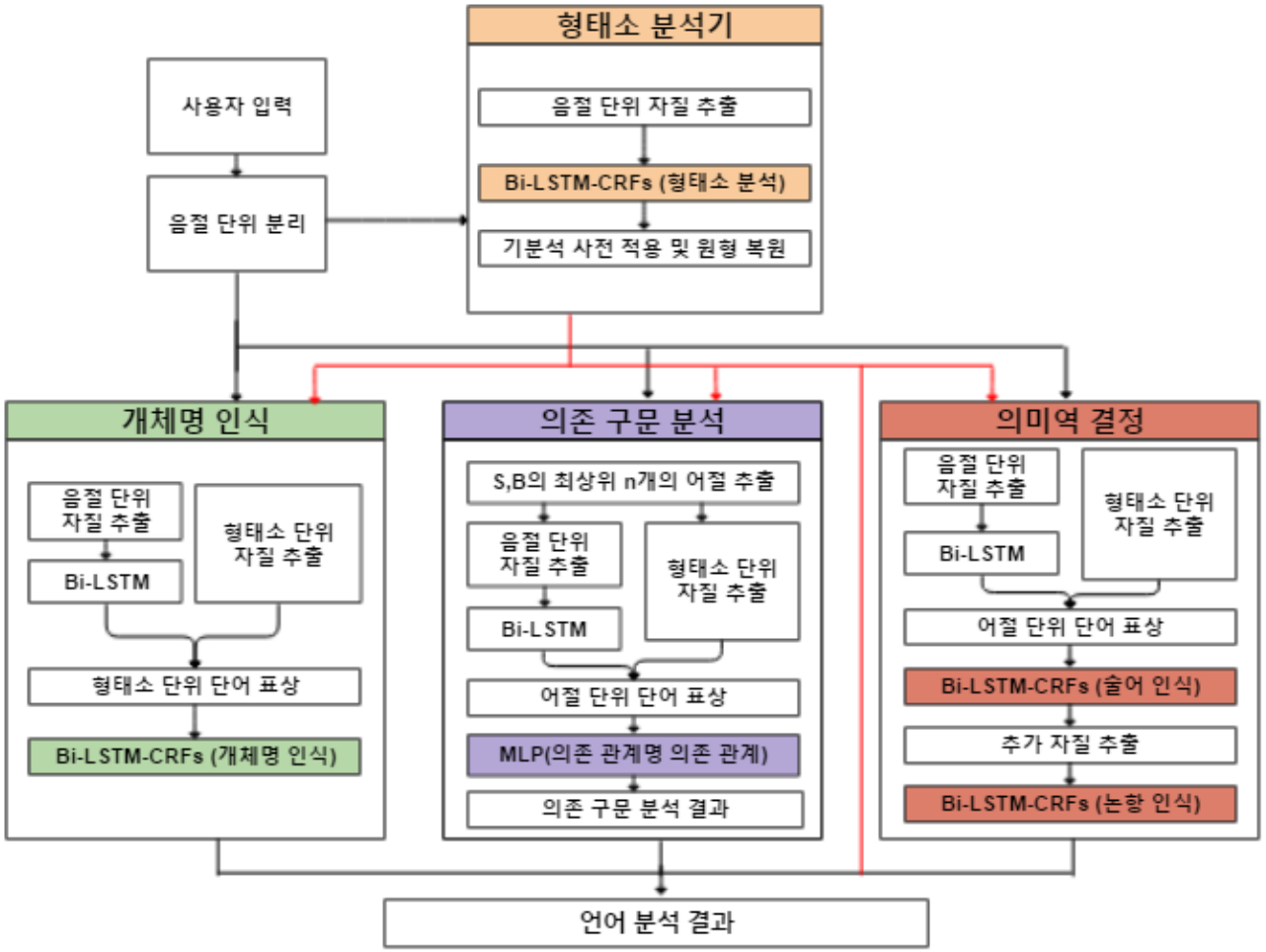
한국어 문장을 효율적이고, 정확한 분석을 지향

특징

- 순차 레이블링에 높은 성능을 내는 Bi-LSTM-CRFs 사용
- 음절 기반 임베딩 벡터로 어절을 표현하는 기법
- 레이블 분포를 이용한 전체적인 분석기 성능 향상

시스템 아키텍처

❖ 언어분석 통합시스템(단추)



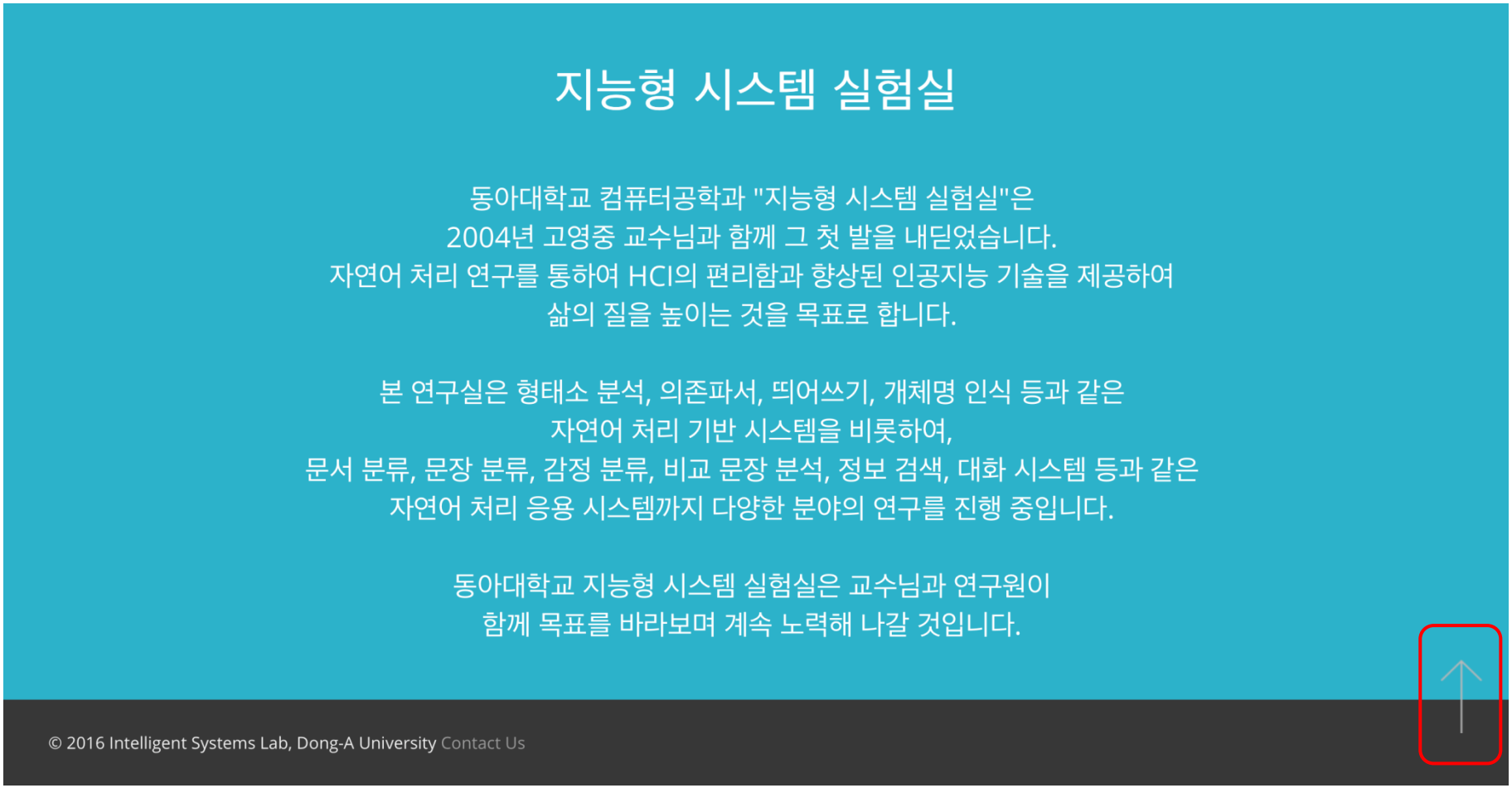
Interface - Main

Interface - Main

입력창 - 분석할 문장을 입력합니다.

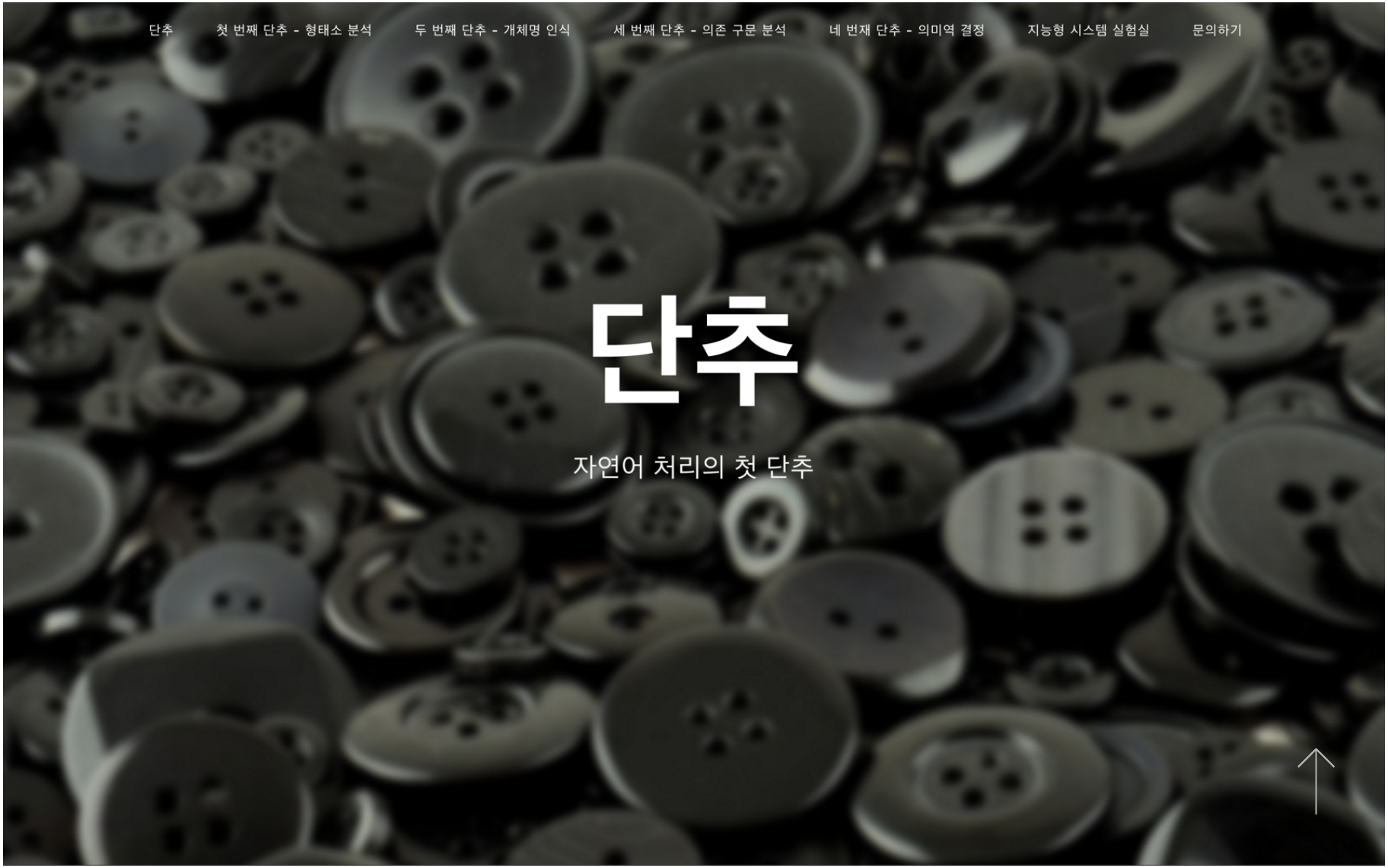
문장을 입력하지 않고 버튼을 누르면 예시 문장이 자동으로 입력됩니다.

Interface - Main

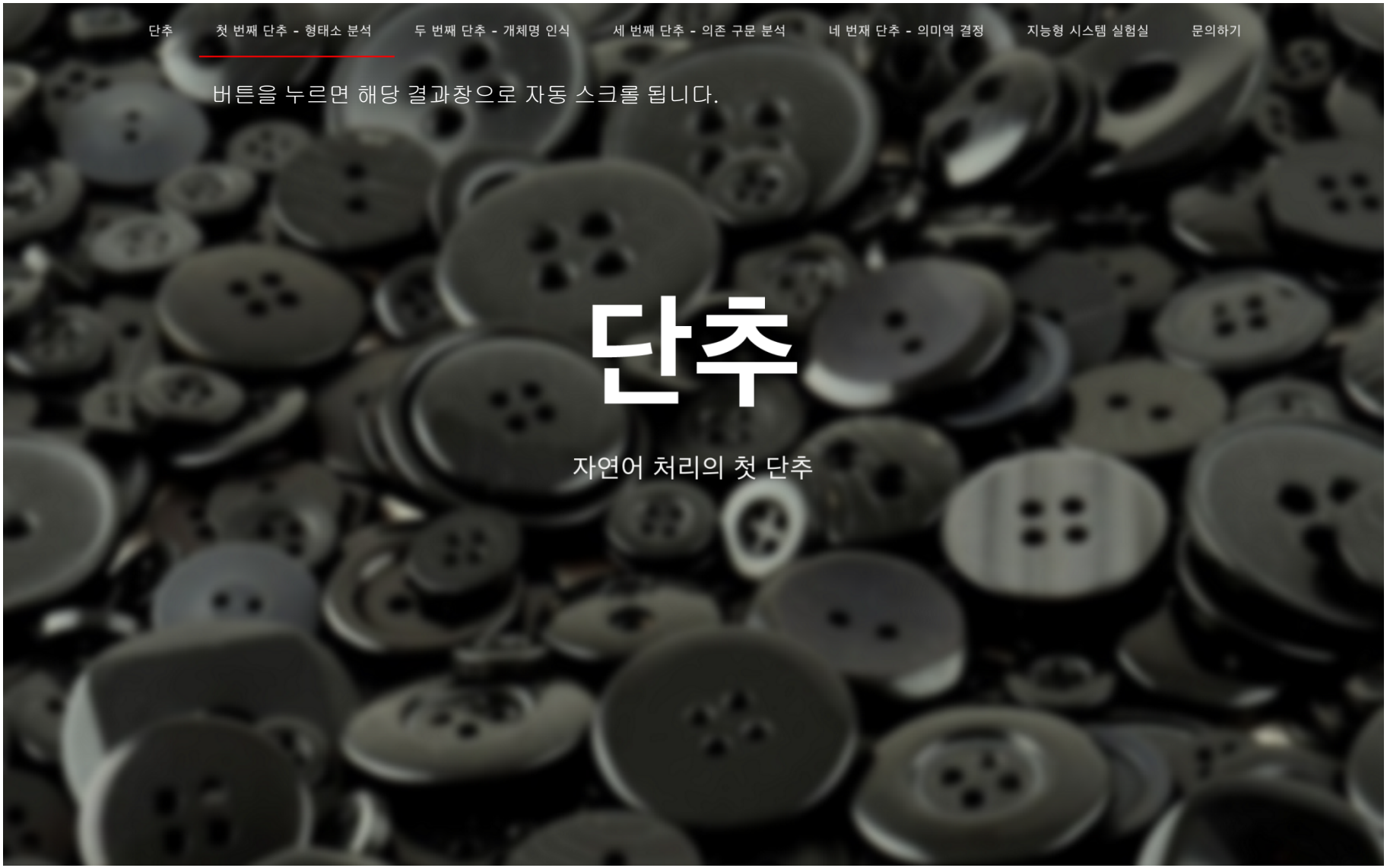


클릭시 페이지 상단으로 이동합니다.

Interface - Main



Interface — POS Tagging



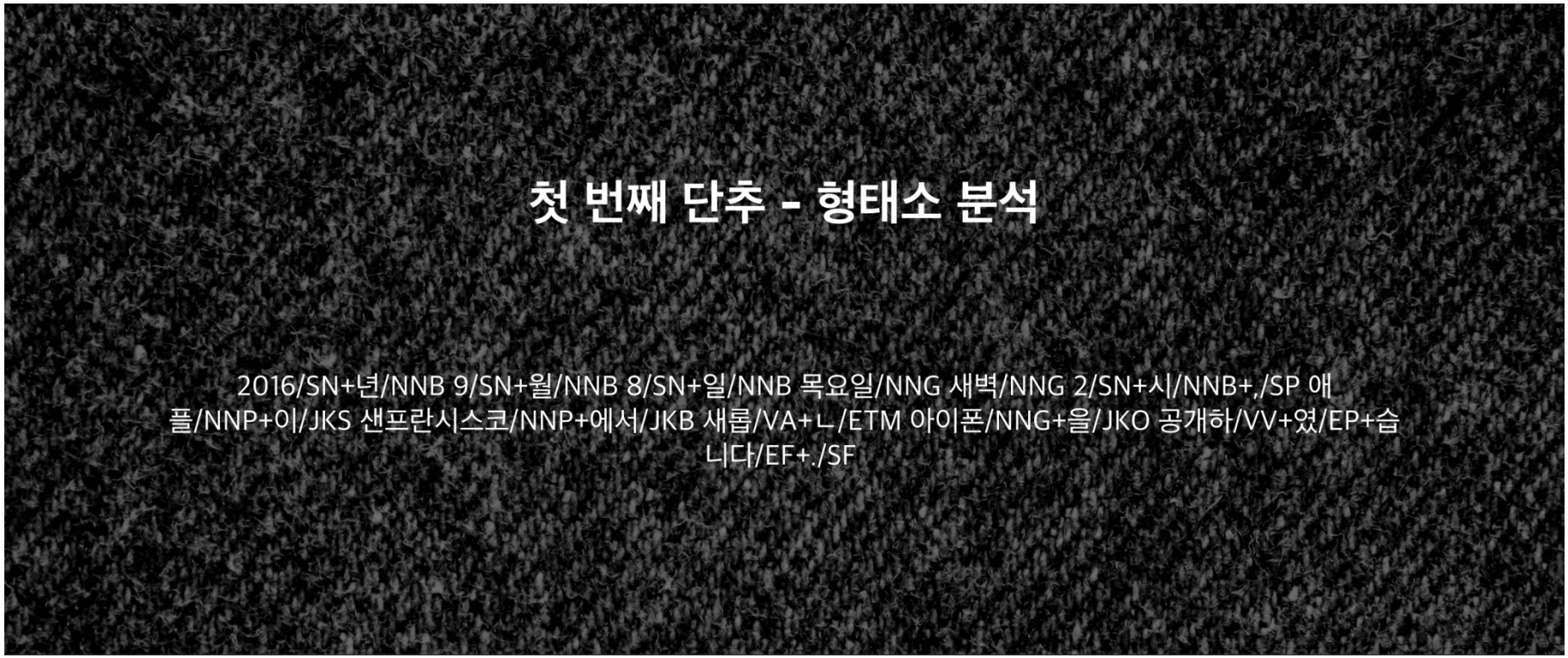
Interface – POS Tagging

입력창 – 이전에 입력한 문장이 입력되어 있습니다.

다른 기능 버튼을 누르면 입력되어 있는 문장으로 처리됩니다.

새로운 문장을 입력할 수도 있습니다.

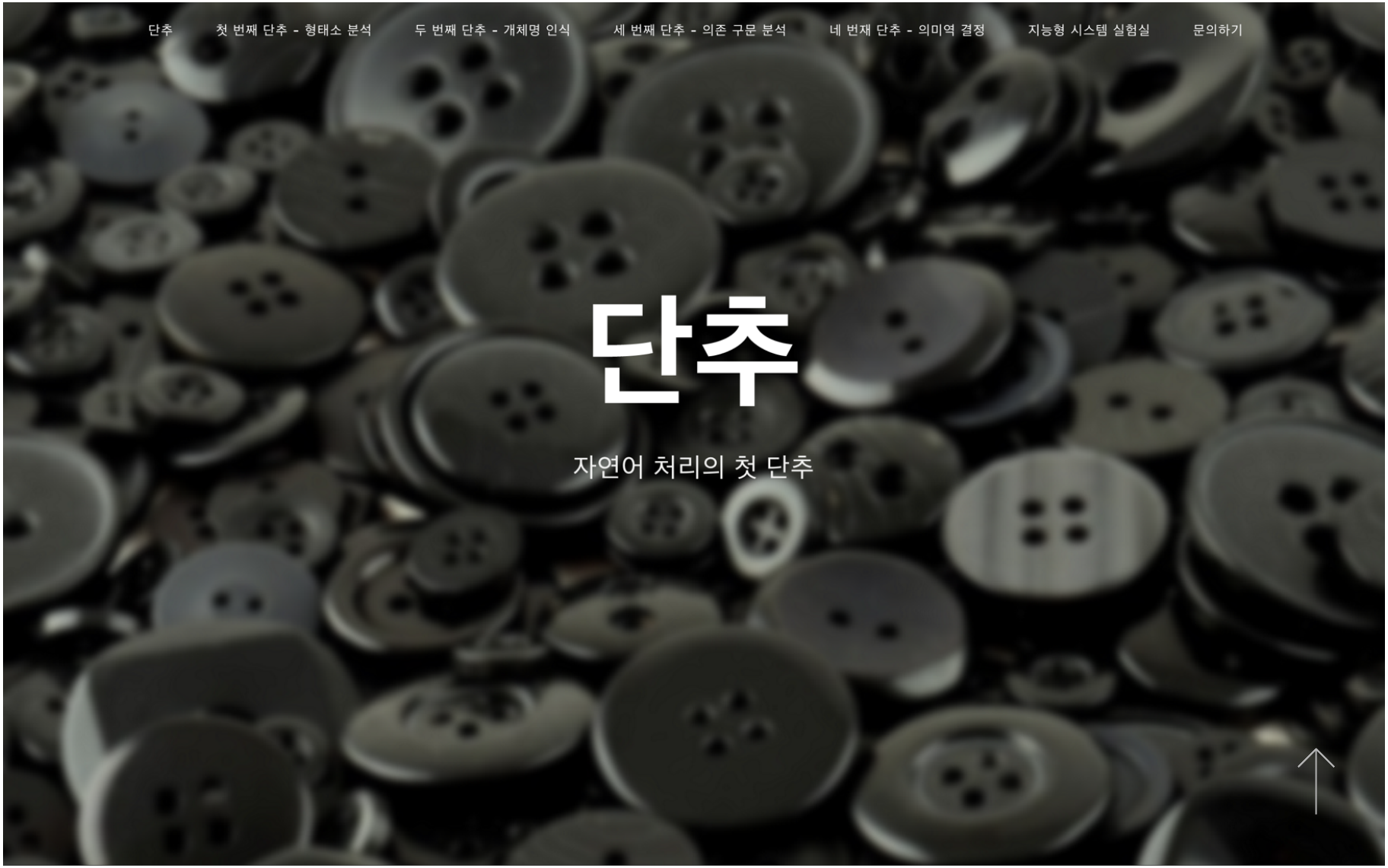
Interface – POS Tagging



결과창 - 형태소 분석 결과를 보여줍니다.

패럴랙스 효과가 적용되어 있습니다.

Interface — POS Tagging



Interface – NER

결과창 – 개체명 인식 결과를 보여줍니다.

개체명 종류 별로 하이라이팅을 해 줍니다.

Interface – DP

세 번째 단추 - 의존 구문 분석

1 4 안재현/NNG NP_OBJ

2 6 연구생/NNG+은/JX NP_SBJ

3 4 가방/NNG+을/JKO NP_OBJ

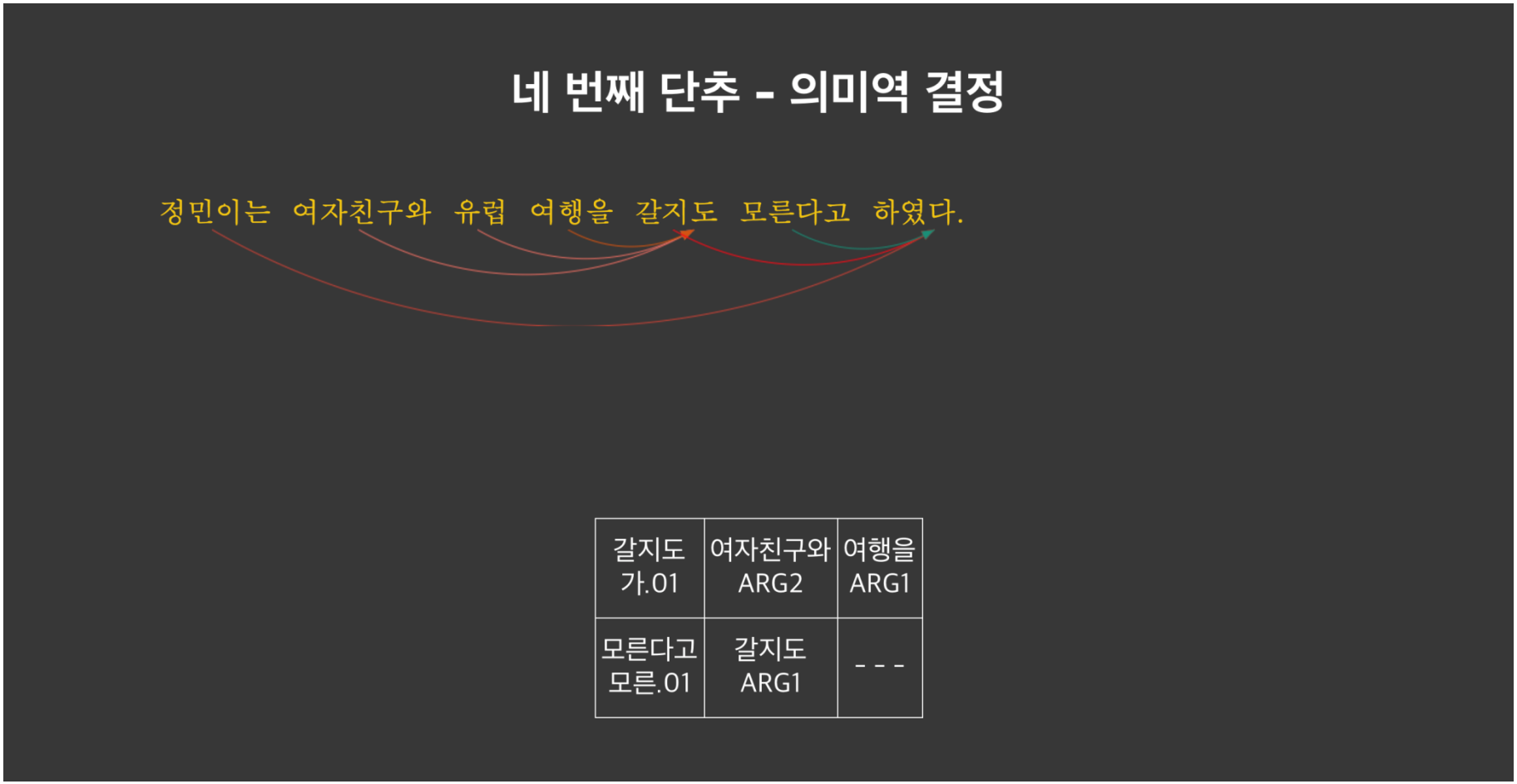
4 6 매/VV+고/EC VP

5 6 학교/NNG+에/JKB NP_AJT

6 0 가/VV+았/EP+다/EF+./SF VP

결과창 – 의존 구문 분석 결과를 화살표로 보이고 하이라이팅을 해 줍니다.

Interface – SRL



결과창 – 의존 구문 분석 결과를 화살표로 표시하고

의미역 결정 결과는 표로 보여줍니다.

언어 분석 기술

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

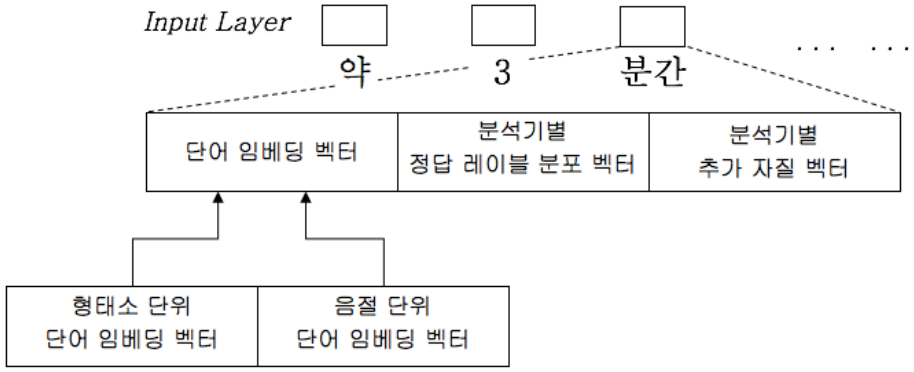
  - 딥 러닝 기반 모델
    - **Bidirectional LSTM CRF**

  - 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 Representations)

- 단어 입베딩 벡터
  - 형태소 단위 임베딩 벡터
  - 음절 단위 임베딩 벡터



[이질적인 단어표상의 연결(Concatenation)]

- 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)

단어 임베딩(Word Embedding)

❖ 단어 임베딩 벡터

- **Skip-gram**을 이용한 학습
- **11.5GB** 뉴스 데이터 이용
  - 전체 약 **22**억 **4,400**만 형태소
  - **Vocabulary Size :** 약 **191**만 형태소

- 단위 별 예시

[표1] 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 **46**차원 **(**품사태그 **46**개**)**
- 한 단어가 각 품사 태그를 정답으로 가질 확률을 벡터로 만들어 사용
  - ✓ 학습 데이터에서 구축

- 품사 분포 벡터 예시

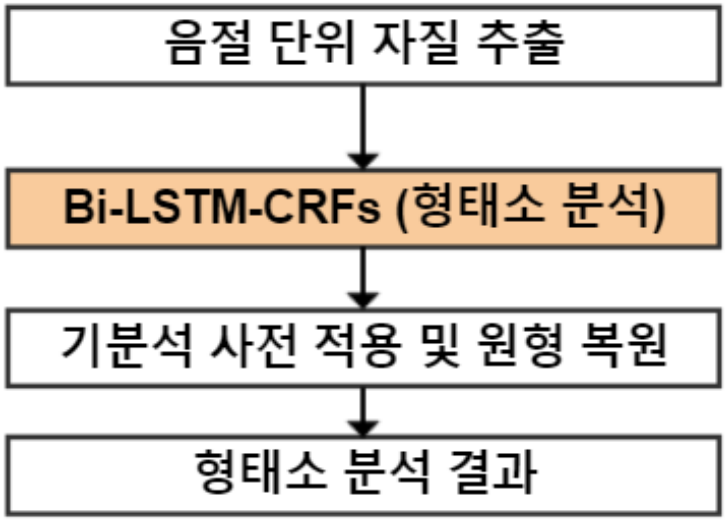
[표2] 음절 단위 품사 분포 벡터 예시

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, ...

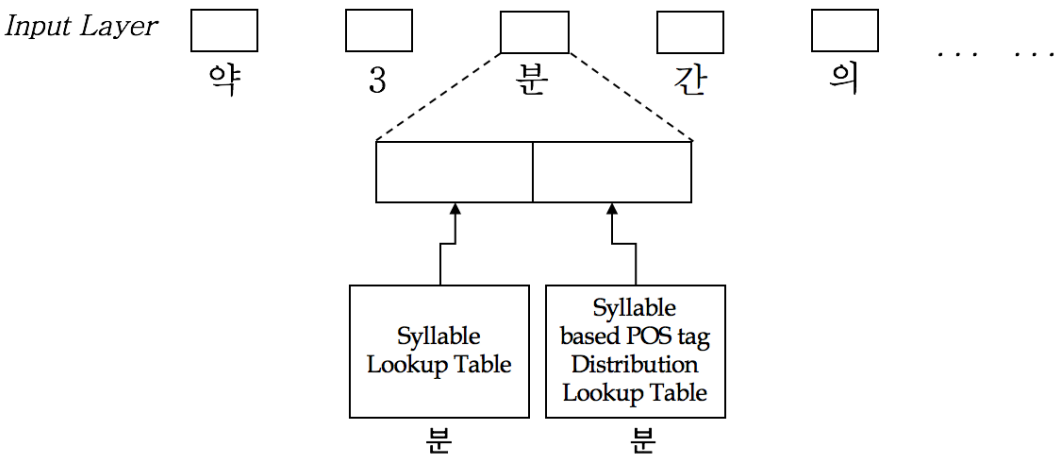
약/MM, 3/SN+분간/NNG+의/JKG, ...

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**)

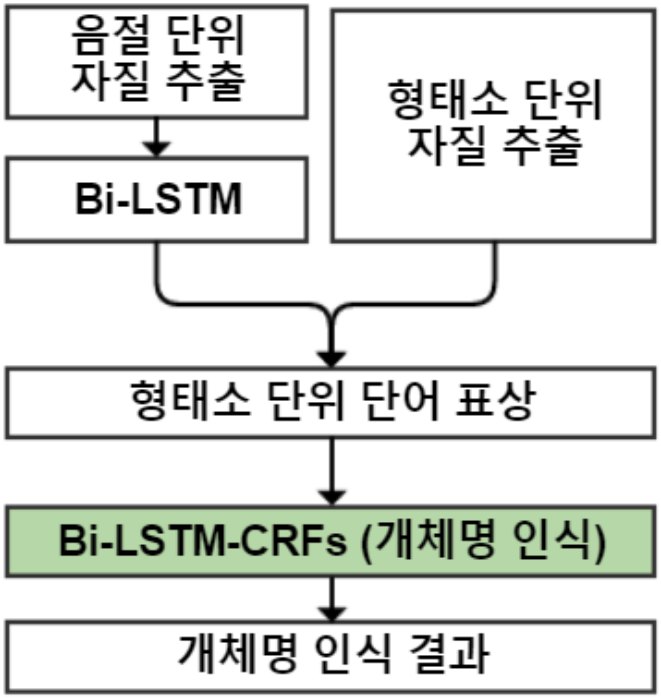


[형태소 분석 입력 단어표상]

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

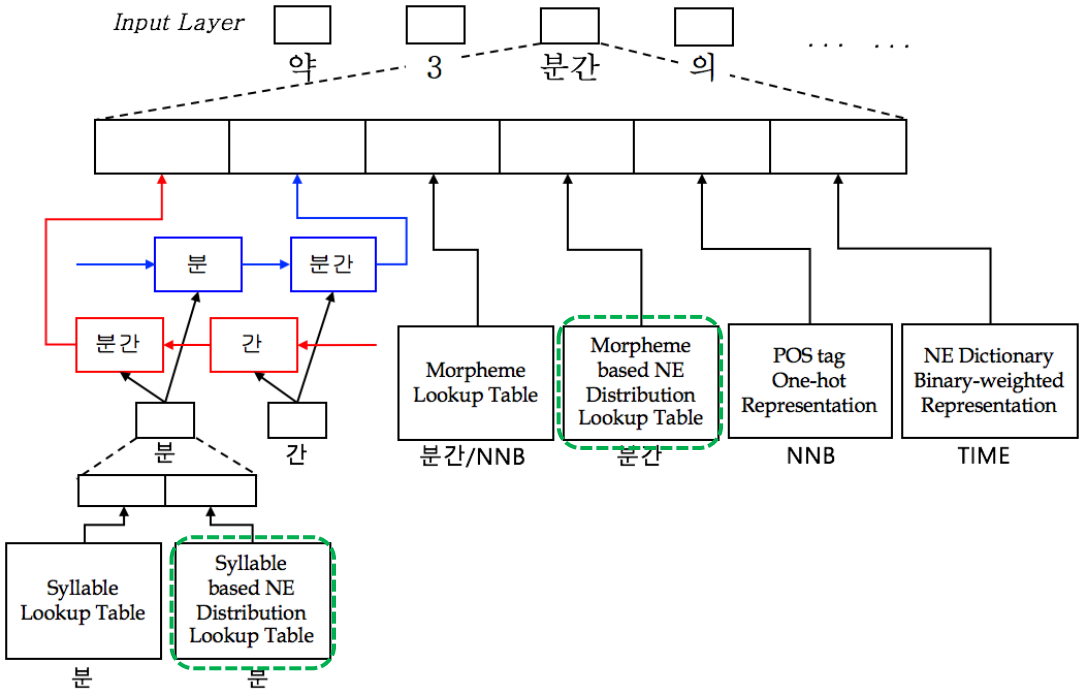
약 3분간의 예열은 엔진을 보호합니다.

개체명 인식기

❖ 개체명 인식

➢ 입력 단어 표상(**Word Representations**)

- 음절 기반 형태소 단위 입력
  - ✓ **Bidirectional LSTM**을 이용한 확장
- 형태소 단위 입력



[개체명 인식 입력 단어표상]

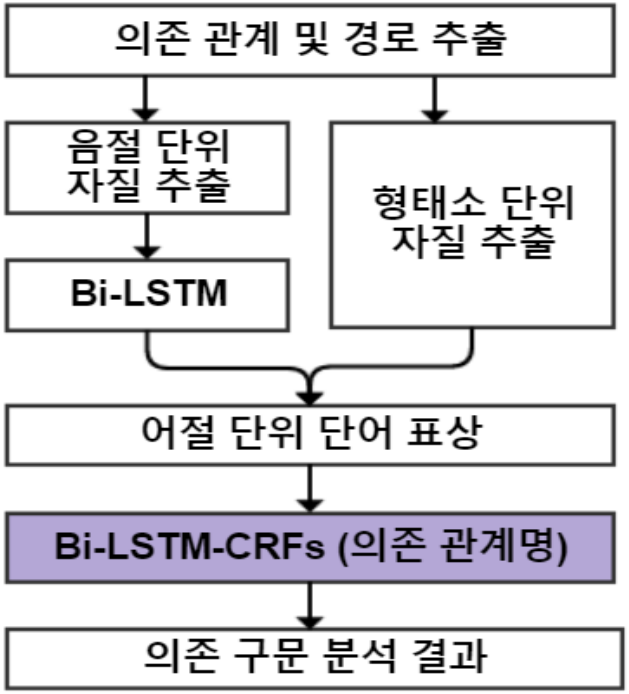
의존 구문 분석기

❖ 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

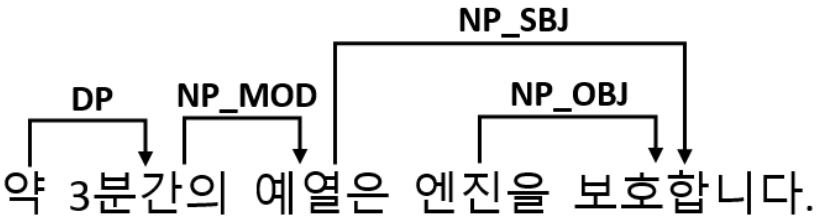


[의존 관계명 분석 모델]

약 → 3분간의 → 예열은 → 보호합니다.
엔진을 → 보호합니다.

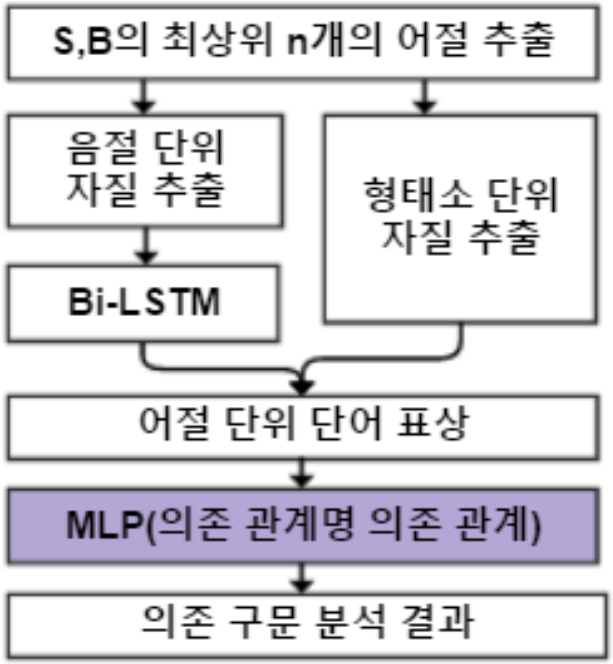
약, 3, 분, 간, 의, ...
약/MM 3/SN+분간/NNG+의/JKG, ...

약, 3분간의, 예열은, 엔진을, 보호합니다.

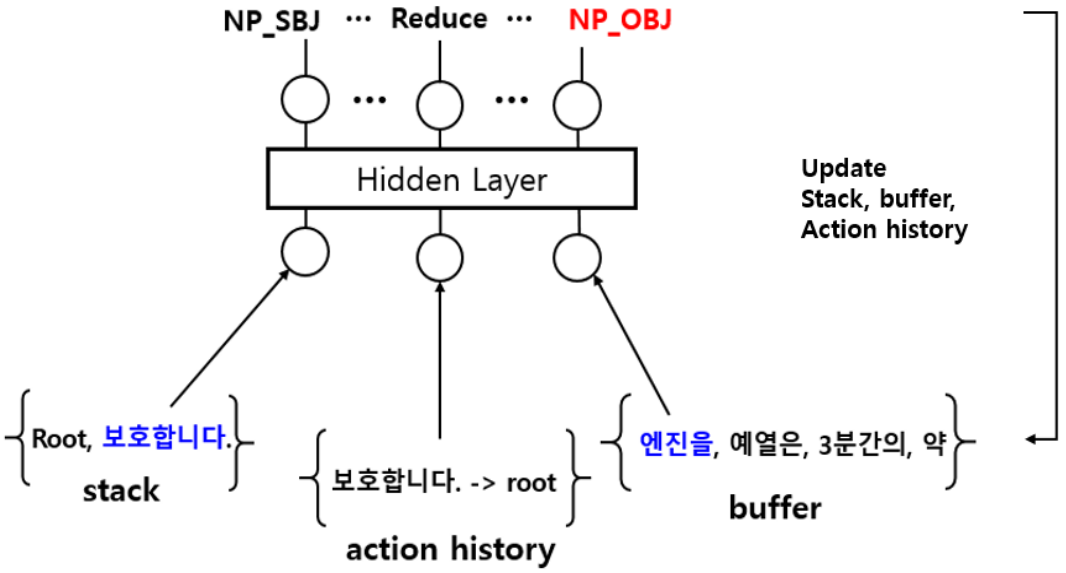


의존 구문 분석기

❖ 의존 관계 및 의존 관계명 분석 모델



[의존관계 및 의존관계명 동시 분석 모델]



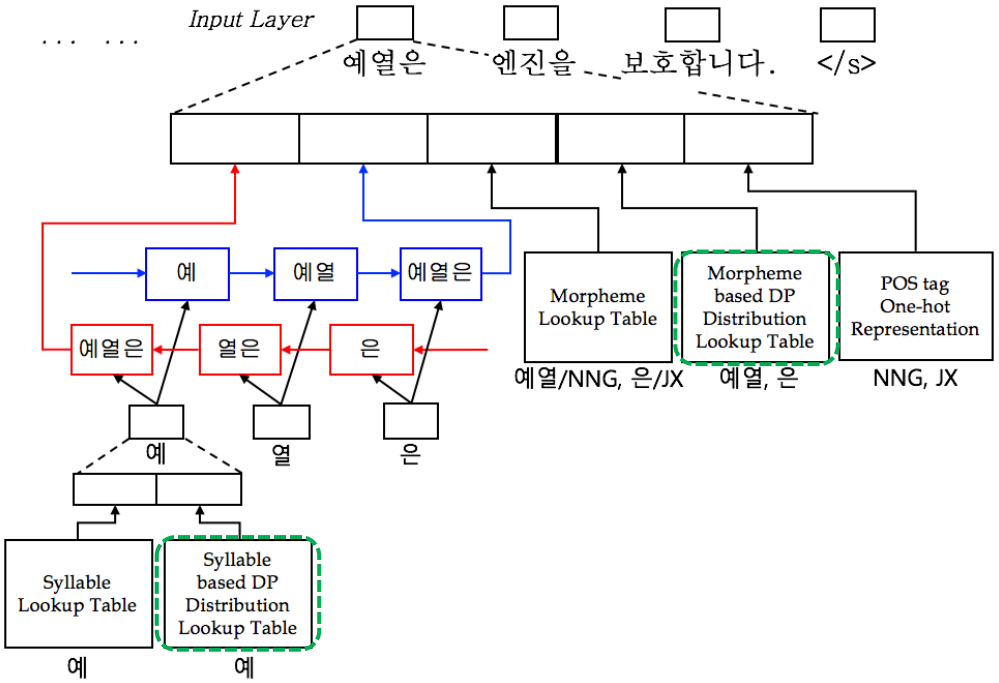
[MLP 기반 컨트롤 네트워크 구성도]

의존 구문 분석기

❖ 의존 구문 분석

➢ 입력 단어 표상(**Word Representations**)

- 음절 기반 어절 단위 입력
  - ✓ **Bidirectional LSTM**을 이용한 확장
- 어절 단위 입력

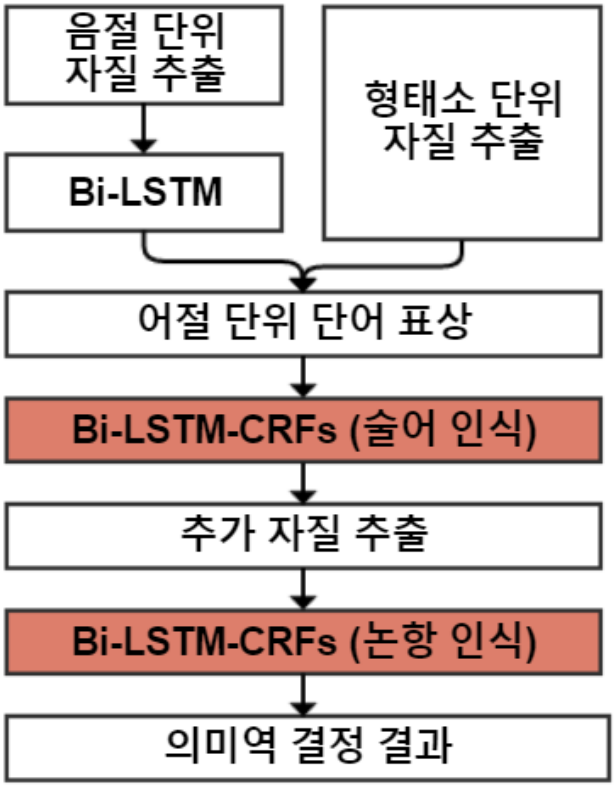


[의존 구문 분석 입력 단어표상]

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

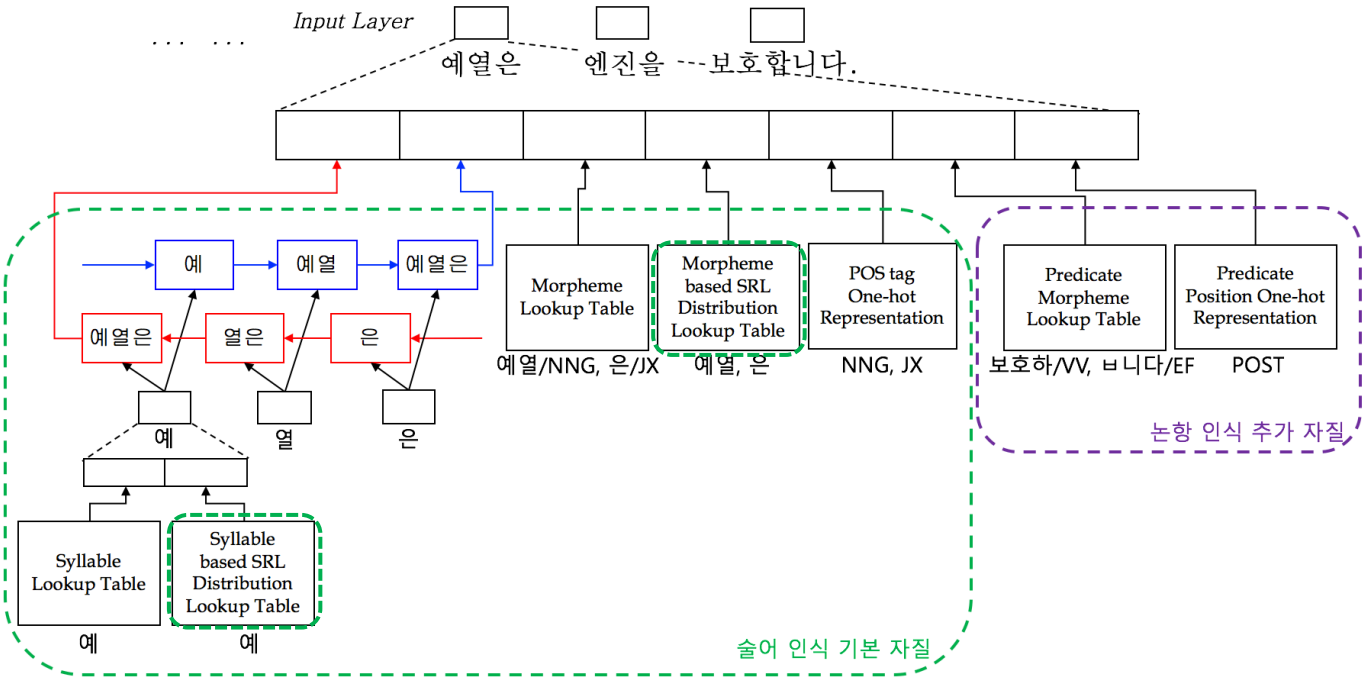
보호하/VV+ㅂ니다/EF

의미역 결정기

❖ 의미역 결정

➢ 입력 단어 표상(**Word Representations**)

- 음절 기반 어절 단위 입력
  - ✓ **Bidirectional LSTM**을 이용한 확장
- 어절 단위 입력



[의미역 결정 입력 단어표상]

성능 평가

❖ 실험 환경(형태소 분석)

➢ 세종 말뭉치

- **8,376** 문장 (학습 **6,701** 문장, 평가 **1,675** 문장)
- **46** 레이블

[표3] 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 (2011)	**99.03**	-
단추 (Ours)	**98.04**	**98.65**

성능 평가

❖ 실험 환경(개체명 인식)

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

[표4] 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 알짬 (2016)	**87.62**	70.79	80.89
KoNER (2016)	85.76	73.83	81.00
단추(Ours)	**87.08**	**76.54**	**82.86**

[표5] 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**

성능 평가

❖ 실험 환경(의존 구문 분석)

- ➢ 세종 말뭉치
  - **59,574** 문장 (학습 **53,757** 문장, 평가 **5,817** 문장)
  - **36** 레이블
- ➢ **[표7]** 의존 구문분석 성능은 포인트 네트워크를 이용한 의존 구문 분석 성능

[표6] 의존 관계명 부착 성능

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

[표7] 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)	**90.69**	**87.5**

성능 평가

❖ 실험 환경(의미역 결정)

➢ Korean PropBank

- **4,853** 문장 (학습 **3,883** 문장, 평가 **970** 문장)
- **23** 레이블

[표8] 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**

결론

❖ 결론

- bi-LSTM-CRFs, 음절 기반 임베딩 벡터
- 향후 자연어 처리연구에 자연어 응용분야의 좋은 토대가 될 것.

Publications

❖ 논문

- 김혜민, 윤정민, 안재현, 배경만, 고영중

  " 품사 분포와 Bidirectional LSTM CRFs를 이용한 음절 단위 형태소 분석기 ",

  제 28회 한글 및 한국어 정보처리 학술대회 (HCLT-2016), pp. 3-8, October 2016

- 김혜민, 양선, 고영중

  " 기분석 어절사전 축소를 통한 한국어 형태소 분석에서의 메모리 및 수행 시간 최적화 ",

  2017 한국소프트웨어종합학술대회

- 유홍연, 고영중 (우수 논문상)

  " 품사 임베딩과 음절 단위 개체명 분포 기반의 Bidirectional LSTM CRFs를 이용한

  개체명 인식 " , 제 28회 한글 및 한국어 정보처리 학술대회 (HCLT-2016), pp. 105-110,

  October 2016

- 유홍연, 고영중

  "Bidirectional LSTM CRF 기반의 개체명 인식을 위한 단어 표상의 확장",

  정보과학회논문지:(KIISE), 제 44권, 제 3호, pp. 306-313, 2017년 3월. (ISSN 2383-6296)

Publications

❖ 논문

- 윤정민, 배경만, 고영중

  " 음절의 의미역 태그 분포를 이용한 Bidirectional LSTM CRFs 기반의 한국어 의미역 결정 " , 제 28회 한글 및 한국어 정보처리 학술대회 (HCLT-2016), pp. 324-329, October 2016

- 윤정민, 고영중

  " 음절 의미역 태그 분포를 이용한 입력 벡터 확장을 통한 Bi-LSTM-CRFs 기반의 의미역 결정 모델 " , 2017 한국인지과학회 연차학술대회, P48, May 2017

- 안재현, 이호경, 고영중

  " 의존 경로와 음절단위 의존 관계명 분포 기반의 Bidirectional-LSTM-CRFs를 이용한 한국어 의존 관계명 레이블링 ", 제 28회 한글 및 한국어 정보처리 학술대회 (HCLT-2016), pp. 14-19, October 2016

- 안재현, 고영중

  " 음절 단위 태그 분포와 멀티 태스크 학습 기반 포인터 네트워크를 이용한 한국어 의존 구문 분석 ", 2017 한국소프트웨어종합학술대회