WiseNLU:
지식처리를 위한 자연어 의미 이해 기술

2015. 8. 21.

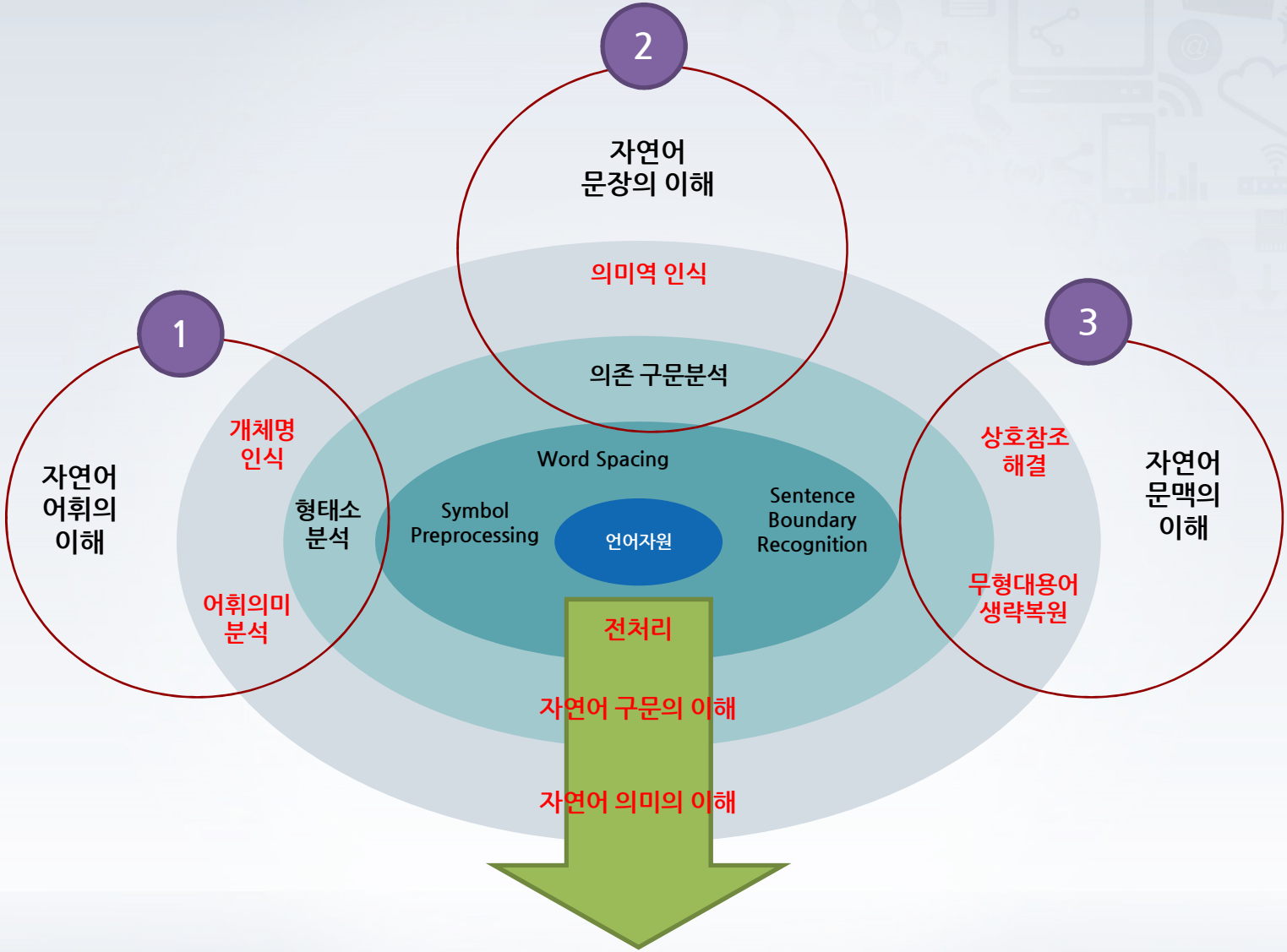
임수종/이충희/임준호/김현기

ETRI 지식마이닝연구실

과제 개념 : WiseQA

복잡한 자연어로 기술된 문제의 의미를 이해하고 정답을 추론하여 생성함

WiseNLU: 개발방향

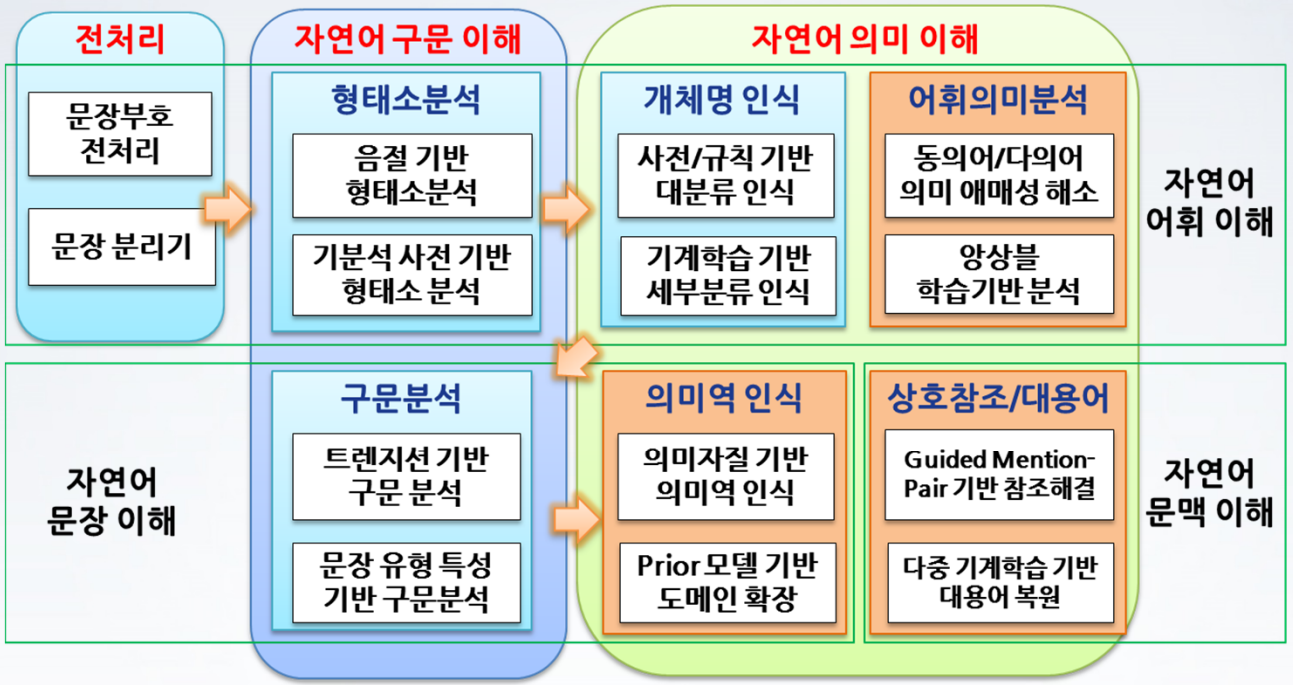


WP2: 지속적 언어지식 학습 연계 성능개선 추진

WiseNLU:연구 목표 및 구성도

언어지능: 세계 최고 수준 자연어 이해 기술 개발 가능성 검증

- 방법론: 인간의 언어이해 방법을 모방한 자연어 이해기술 설계 및 방법론 정립
  1. 파이프라인: 어휘분석 → 어휘의미분석 → 구문분석 → 의미역인식 → 문맥인식
  2. 하이브리드: 빠르고 명확한 분석 방법(사전+패턴) + 의미 제약/추론 방법[어휘의미망+기계학습)



WiseNLU: Example

소피스트란 그리스어로 지혜로운 자 또는 지혜를 만들어내는 사람이라는 뜻으로, BC 5~4세기의 그리스의 철학자들을 말한다. 이들은 아테네 사람들을 대상으로 하였고, 수사학과 웅변술을 가르쳤다.

형태소 분석

소피스트/NNG+란/JX 그리스/NNP+어/XSN+로/JKB 지혜롭/VA+ㄴ/ETM 자/NNB 또는/MAG 지혜/NNG+를/JKO 만들/VV+어/EC+내/VX+는/ETM 사람/NNG+이/VCP+라는/ETM 뜻/NNG+으로/JKB+,/SP BC/SL 5/SN+~/SO+4/SN+세기/NNP+의/JKG 그리스/NNP+의/JKG 철학/NNG+자/XSN+들/XSN+을/JKO 말/NNG+하/XSV+s다/EF+./SF 이/NP+들/XSN+은/JX 아테네/NNP 사람/NNG+들/XSN+을/JKO 대상/NGG+으로/JKB 하/VV+았/EP+고//EC+,/SP 수사/NNG+학/XSN+과/JC 웅변/NNG+술/XSN+을/JKO 가르치/VV+었/EP+다/EF+./SF

개체명 인식

<CV_OCCUPATION:소피스트/NNG>+란/JX <CV_LANGUATE:그리스/NNP+어/XSN>+로/JKB 지혜롭/VA+ㄴ/ETM 자/NNB 또는/MAG 지혜/NNG+를/JKO 만들/VV+어/EC+내/VX+는/ETM 사람/NNG+이/VCP+라는/ETM 뜻/NNG+으로/JKB+,/SP <DT_DURATION:BC/SL 5/SN+~/SO+4/SN+세기/NNP>+의/JKG <LCP_COUNTRY:그리스/NNP>+의/JKG 철학/NNG+자/XSN+들/XSN+을/JKO 말/NNG+하/XSV+s다/EF+./SF 이/NP+들/XSN+은/JX <LCP_CAPITALCITY:아테네/NNP> 사람/NNG+들/XSN+을/JKO 대상/NGG+으로/JKB 하/VV+았/EP+고//EC+,/SP <FD_ART:수사/NNG+학/XSN>+과/JC <FD_ART:웅변/NNG+술/XSN>+을/JKO 가르치/VV+었/EP+다/EF+./SF

어휘의미분석

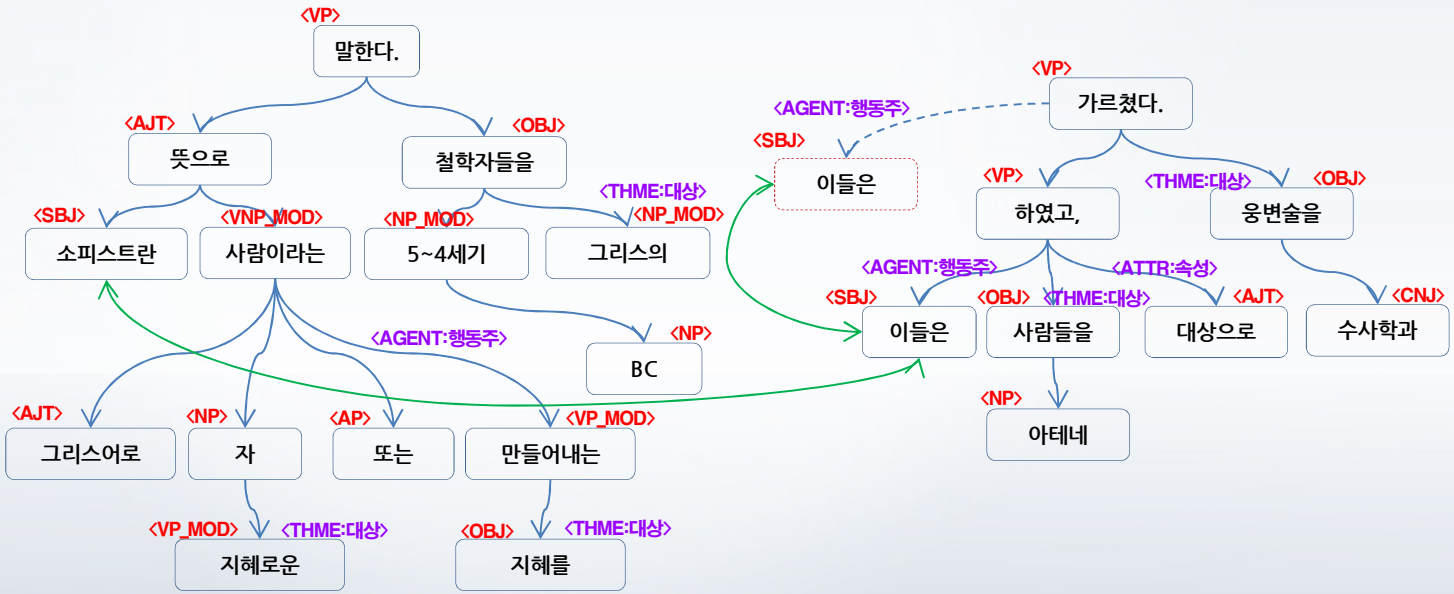
소피스트란 그리스어로 지혜로운 자__18_0000/NNB 또는 지혜__02_0001/NNG+를 만들__00_0101/VV+어내는 사람__00_0001/NNG+이라는 뜻__00_0002/NNG+으로, BC 5~4+세기__03_0002/NNG+의 그리스__02_0000/NNP+의 철학자들을 말하__00_0101/VV+ㄴ다.

의존구문분석

의미역 인식

상호참조해결

무형대용어 생략복원



WiseQA 적용

- 헨리 필립 호프가 블루 호프를 구매한 해는?
  - 정답 후보 문장
    - 은행가 헨리 필립 호프는 블루 호프를 1830년에 구매했다.
    - 헨리 필립 호프는 블루 호프를 이듬해에 샀다.
    - 1830년 헨리 필립 호프가 구매한 블루 호프는 …
    - 블루 호프는 헨리 필립 호프에게 1830년에 팔렸다.
    - 보석상 에리아손은 블루 호프를 헨리 필립 호프에게 1830년에 팔았다.
    - 헨리 필립 호프는 1900년에 뉴욕의 거래상에게 블루 호프를 팔았다.
    - 1830년 헨리 필립 호프는 런던에서 이 다이아몬드를 구입했다.
    - 헨리 필립 호프는 블루 호프를 70년간 소유하였는데, 1830년에 사들였다.

음절 학습 기반 형태소 분석 기술

연구목표 및 성과

- 사전과 음절학습을 통합한 형태소 분석기
  * 기분석 사전 270만건+음절학습 111만 어절
- 다국어 언어 확장을 위한 형태소 분석 방법
  * SVM 기반 언어 독립적인 음절 학습 적용
- 한국어 형태소 태그셋 국내 표준 채택

엔겔키퍼비교수는 엑스선은 파장이 0.01 나노미터이며,…

1. 음절 학습 기반 형태소 분석

엔겔키퍼비/NNP+교수/NNG+는/JX 엑스선/NNG+은/JX 파장/NNG+이/JKS 0.01/SN 나노/NNG+미터/NNB+이/VCP+며/EC+,/SP …

문맥 정보 기반 음절 학습을 통한 신조어 인식

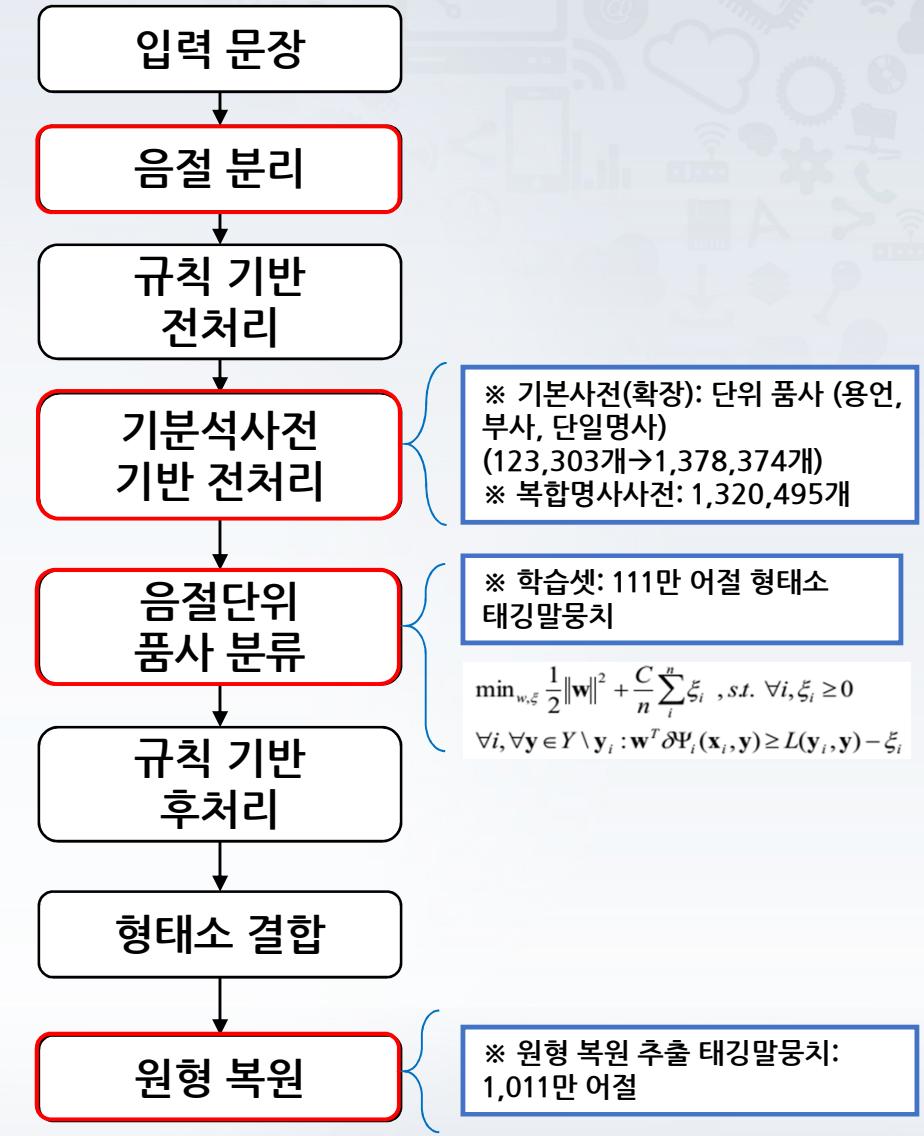
기분석 사전을 통한 음절학습 오류 수정

2. 기분석 사전 기반 형태

엔겔키퍼비/NNP+교수/NNG+는/JX 엑스선/NNG+은/JX 파장/NNG+이/JKS 0.01/SN 나노미터/NNB+이/VCP+며/EC+,/SP …

* NNP:고유명사, NNG:일반명사, NNB:의존명사

〈형태소 분석 기술 〉

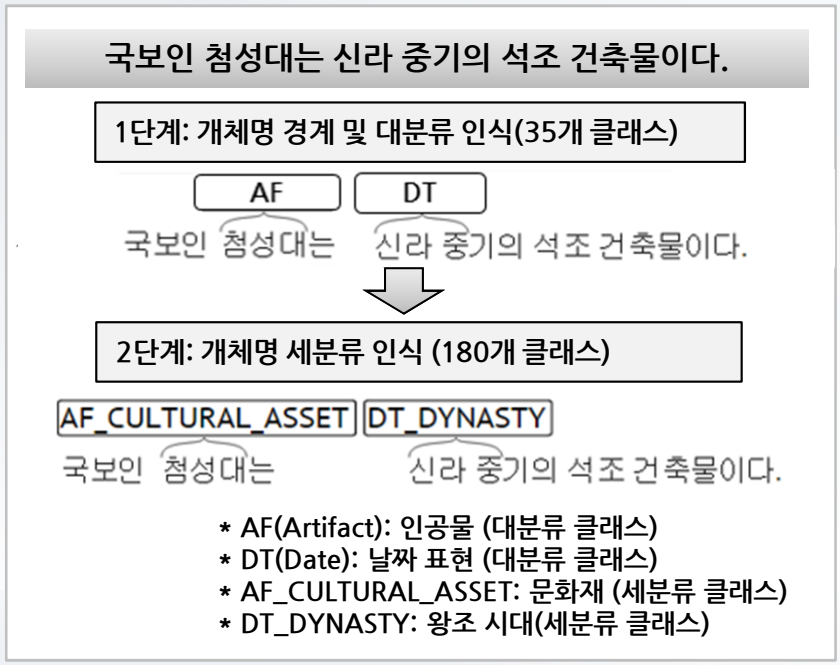


〈형태소 분석 기술 구성도〉

세부분류 개체명 인식 기술

연구목표 및 성과

- 자연어 질의응답을 위한 개체명 인식 기술 개발
  * 2단계, 180개 세부분류 인식
  * 개체명 사전 307만건+학습셋 956만 문장
- 구와 절 형태의 고난이도 개체명 인식 기술 개발
  * 예: 영화 → '성실한 나라의 엘리스'
- 한국어 개체명 분류체계 정립 및 국내 표준화



- Q: <PS_NAME:헨리 필립 호프>가 <MT_ROCK:블루 호프>를 구매한 해는?
- 은행가 <PS_NAME:헨리 필립 호프>는 <MT_ROCK:블루 호프>를 <DT_YEAR:1830년>에 구매했다.
- <PS_NAME:헨리 필립 호프> 는 <MT_ROCK:블루 호프> 를 이듬해에 샀다.
- <MT_ROCK:블루 호프> 는 <PS_NAME:헨리 필립 호프> 에게 <DT_YEAR:1830년> 에 팔렸다.
- 보석상 에리아손은 <MT_ROCK:블루 호프> 를 <PS_NAME:헨리 필립 호프> 에게 <DT_YEAR:1830년> 에 팔았다.
- <PS_NAME:헨리 필립 호프> 는 <DT_YEAR:1900년>에 <LCP_CITY:뉴욕>의 거래상에게 <MT_ROCK:블루 호프> 를 팔았다.
- <DT_YEAR:1830년> <PS_NAME:헨리 필립 호프> 는 <LCP_CITY:런던>에서 이 다이아몬드를 구입했다.
- <PS_NAME:헨리 필립 호프> 는 <MT_ROCK:블루 호프> 를 <DT_DURATION:70년간> 소유하였는데, <DT_YEAR:1830년> 에 사들였다.

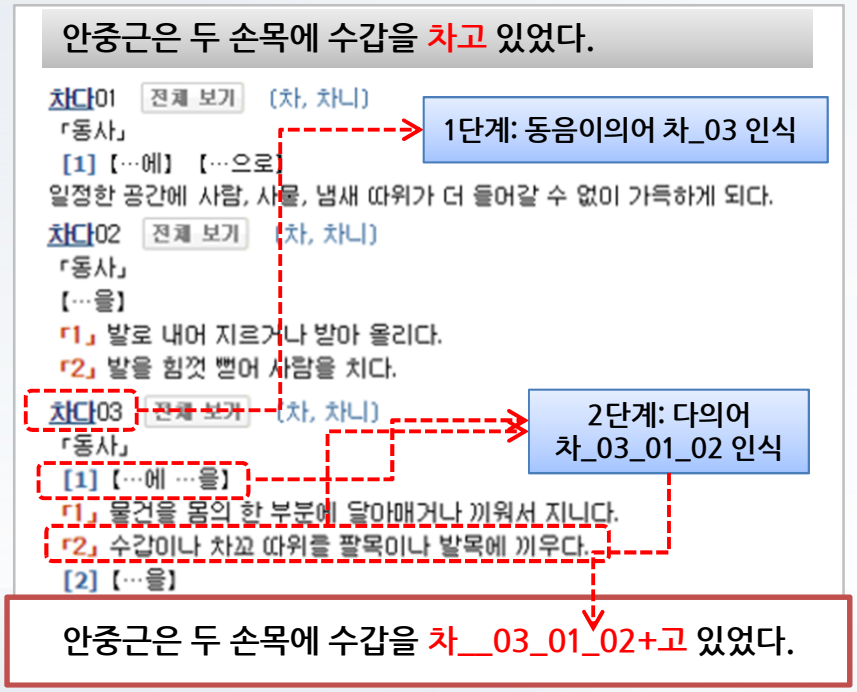
다의어 수준 어휘 의미분석 기술

연구목표

- 의미이해를 위한 동형이의어 및 다의어 분석 기술 개발
  - * 동형이의어 빈도: 9.7% (표준국어대사전)
  - * 다의어 빈도: 12% (표준국어대사전)
- 고성능 어휘 의미분석을 위한 결합 방법론 연구

주요성과

- 동음이의어와 다의어 분석을 순차적으로 분석하는 2단계 어휘의미 분석 기술 개발
  - * 동음이의어 학습셋 818만건 + 다의어 학습셋 377만건
- 다양한 의미분석 모델을 결합한 앙상블 학습 기반 어휘의미분석 방법 확립



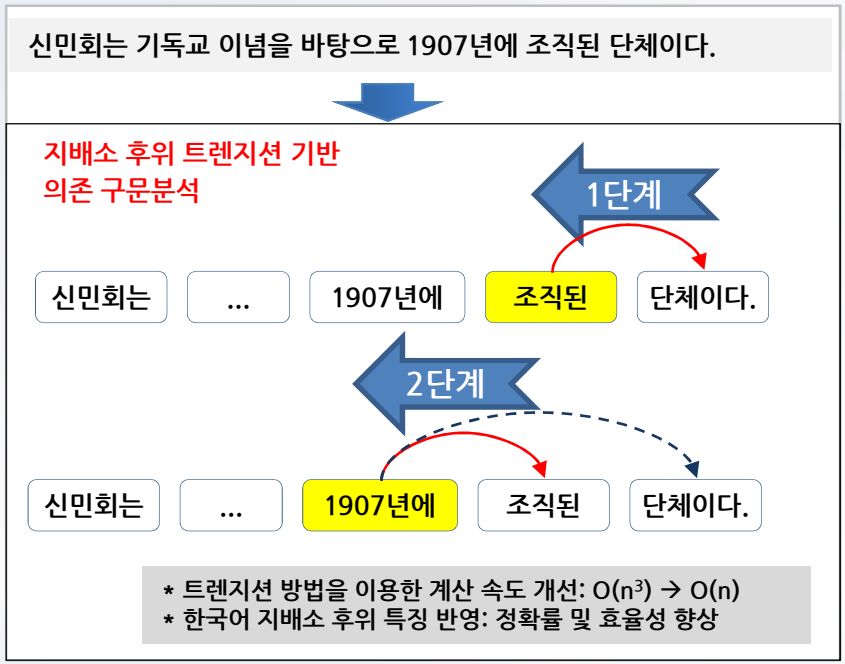
〈어휘의미 분석 기술 연구목표〉

- Q: 헨리 필립 호프가 블루 호프를 구매한 〈해:010002〉는? 〈Q_Focus: temp〉
- 헨리 필립 호프는 블루 호프를 이듬해에 샀다 〈사:000100〉.
- 블루 호프는 헨리 필립 호프에게 1830년에 팔렸다.〈팔리:000001〉
- 보석상 에리아손은 블루 호프를 헨리 필립 호프에게 1830년에 팔았다〈팔:000101〉.
- 헨리 필립 호프는 1900년에 뉴욕의 거래상에게 블루 호프를 팔았다〈팔:000101〉.

의존 구문분석 기술

연구목표

- **지배소 후위 트렌지션** 기반 의존구문분석 개발
  - * 250여종 자질 개발 및 최적화 기계학습 기술 개발
  - * **문장 부호 및 문장 유형 특성**을 반영한 성능 개선
  - * **국내 최고 정확률 92.5%(세종), 93.0%(GS) 달성**
- 위키피디아와 다양한 문장유형의 의존구문 분석
- 기계학습과 규칙을 혼합한 하이브리드 분석



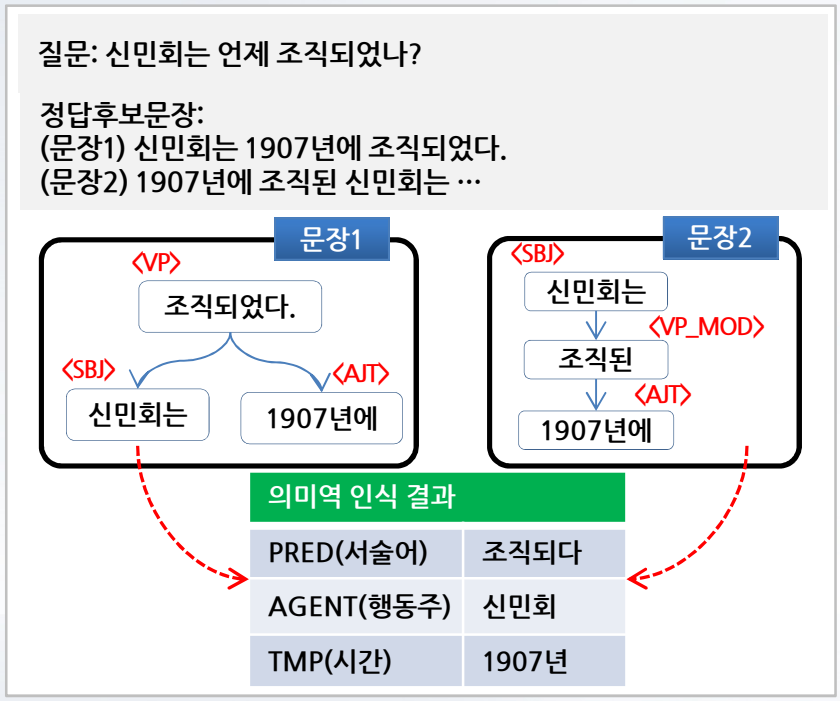
- Q:헨리 필립 호프가 블루 호프를 구매한 해는?
  구매하다(헨리 필립 호프:SBJ, 블루 호프:OBJ, 해:AJT_temp → Q_focus)
- 은행가 헨리 필립 호프는 블루 호프를 1830년에 구매했다.
  구매하다(헨리 필립 호프:SBJ, 블루 호프: OBJ, 1830년:AJT_temp)
- 헨리 필립 호프는 블루 호프를 이듬해에 샀다.
  사다$_{000100}$(헨리 필립 호프:SBJ, 블루 호프: OBJ, 이듬해:AJT_temp)
  - 다의어 수준 어휘의미분석
  - 유의어 정보(구매하다 == 사다$_{000100}$)
  - 시간정보 정규화(이듬해 == 1830년)
- 1830년 헨리 필립 호프가 구매한 블루 호프는…
  구매하다(헨리 필립 호프:SBJ, 블루 호프:VP_MOD, …)
- 장 밥티스트는 블루 호프를 헨리 필립 호프에게 1830년에 팔았다.
  팔다(장 밥티스트:SBJ, 블루 호프:OBJ, 헨리 필립 호프:AJT, 1830년:AJT_temp)

* UAS: Unlabeled Attachment Score

의미역 인식 기술

연구목표

- 문장 표현의 의미애매성 해소를 위한 의미역 인식 기술 개발
  * Sequence Labeling 및 의미 자질 활용
- 도메인 확장을 위한 Prior model 기반 DA 방법론 확립
- 한국어 의미역 및 태깅 말뭉치 구축 방법론 정립
  * 의미역 개수: 23개 (핵심격 5개, 부가격 18개)



- Q:헨리 필립 호프가 블루 호프를 구매한 해는?
  구매하다(헨리 필립 호프:A0-buyer, 블루 호프:A1-thing bought, 해:AM-TMP → Q_focus)
- 1830년 헨리 필립 호프가 구매한 블루 호프는…
  구매하다(헨리 필립 호프:A0-buyer, 블루 호프:A1-thing bought, 1830년:AM-TMP)
- 보석상 에리아손은 블루 호프를 헨리 필립 호프에게 1830년에 팔았다.
  팔다(에리아손:A0-seller, 블루 호프:A1-thing sold, 헨리 필립 호프:A2-buyer, 1830년:AM_TMP)
  - FrameSet (구매하다 ←→팔다)
- 헨리 필립 호프는 1900년에 뉴욕의 거래상에게 블루 호프를 팔았다.
  팔다(헬리 필립 호프:A0-seller, 블루 호프:A1-thing sold, 뉴욕의 거래상:A2-buyer, 1900년:AM_TMP)
- 블루 호프는 헨리 필립 호프에게 1830년에 팔렸다.
  팔다(블루 호프:A1-thing sold, 헨리 필립 호프:A2-buyer, 1830년:AM_TMP)
  - 사동-피동 관계 (팔리다 ←→ 팔다 ←→ 구매하다)

* DA: Domain Adaptation

상호참조 해결 기술

연구목표

- 규칙과 통계를 결합한 한국어 상호참조해결 기술 개발
  * 상호참조 사용 빈도: 문장 당 2.8회 (위키백과 889 문장 분석 결과)
- 한국어 상호참조해결 기술 정립 및 표준화

질문: 1907년에 안창호가 설립한 조직은?

정답후보문장:
신민회는 기독교 이념을 바탕으로 조직된 비밀결사단체이다. 이 단체는 1907년 4월에 안창호의 발기에 의해서 창립되었다.

정답: 이 단체?

상호참조해결 결과
신민회는 기독교 이념을 바탕으로 조직된 비밀결사단체이다. <이 단체:신민회>는 1907년 4월에 안창호의 발기에 의해서 창립되었다.

정답: 신민회

주요성과

- 국내 최초 규칙/통계 결합 방법 확립
  * Deep Learning 기반 Guided Mention-Pair 모델 개발
  * 세계 최고 정확률 69.6% 달성 (IBM:63.4%)
- 국내 표준화를 위한 한국어 상호참조해결 말뭉치 및 태깅가이드 구축

- Q:헨리 필립 호프가 블루 호프를 구매한 해는?
  구매하다(헨리 필립 호프:A0-buyer, 블루 호프:A1-thing bought, 해:AM-TMP → Q_focus)

- 1830년 헨리 필립 호프는 런던에서 이 다이아몬드를 구입했다.
  구입하다(헨리 필립 호프:A0-buyer, 이 다이아몬드:A1-thing bought, 1830년:AM-TMP, …)

  - 어휘의미정보, 개체명 인식 기반 상호참조해결

무형대용어 생략 복원 기술

연구목표

- 한국어 필수격 무형 대용어 복원 기술 개발
  * 무형 대용어 사용 빈도: 문장 당 0.92회
  (위키피디아 3,000 문장 분석 결과)
- 학습데이터 구축 툴 및 시각화 모듈 개발
- 한국어 무형대용어 기술 정립 및 표준화

질문: 신민회가 정주에 설립한 학교는?

정답후보문장:
신민회는 기독교 이념을 바탕으로 조직된 비밀결사단체이다.
민족 교육을 추진하고자 평양에 대성학교와 정주의 오산학교를 설립하였다.

'정주의 오산학교를 설립하였다': 주어 생략으로 정답 추론 불가

무형대용어 복원 결과
신민회는 기독교 이념을 바탕으로 조직된 비밀결사단체이다.
〈신민회는〉 민족 교육을 추진하고자 평양에 대성학교와 정주의 오산학교를 설립하였다.

'신민회' 주어 복원으로 인해 정답 '오산학교' 추론 가능

주요성과

- 3단계 다중 기계학습 기반 생략 복원 기술 개발
  * 각 단계별 최적 기계학습 방법 적용
  * 세계최고 성능: 69.2% (일본어: 42.7)
- 한국어 무형대용어 태그셋 정립 및 국내 표준화를 위한 말뭉치 구축(위키피디아 3천 문장)

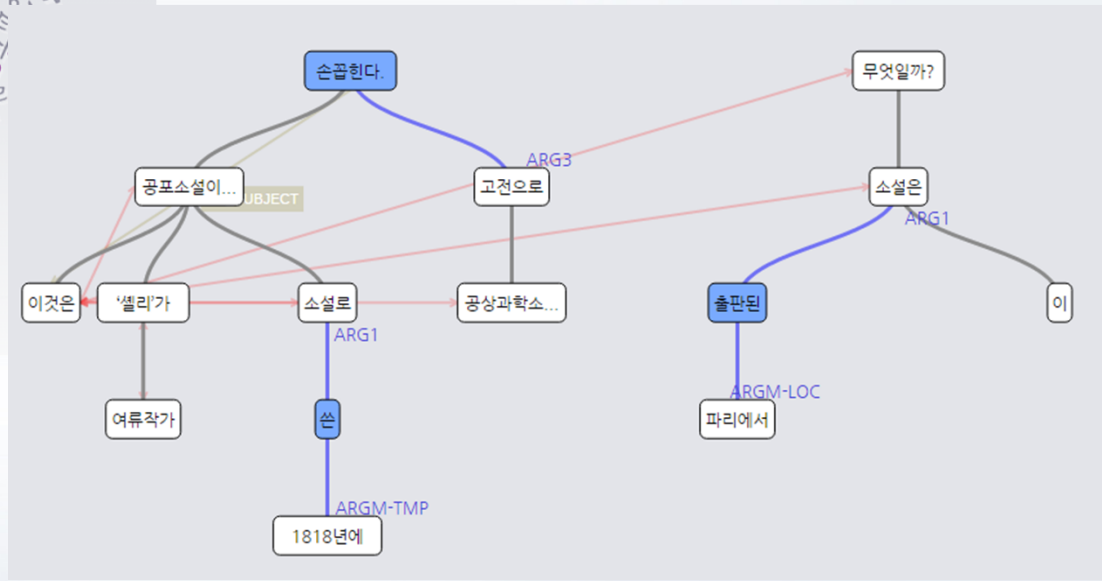
- Q:헨리 필립 호프가 블루 호프를 구매한 해는?
  구매하다(헨리 필립 호프:A0-buyer, 블루 호프:A1-thing bought, 해:AM-TMP → Q_focus)

- 헨리 필립 호프는 블루 호프를 70년간 소유하였는데, 1830년에 사들였다.
  소유하다(헨리 필립 호프:SBJ, 블루 호프:OBJ, 70년간:AJT)
  사들이다(1830년:AJT)

  - 생략된 필수격(주격, 목적격) 복원
  - 표제어 복원

WiseNLU : 데모

- ETRI 언어분석 워크벤치
  - 이것은 여류작가 '셸리'가 1818년에 쓴 소설로 공포소설이면서 공상과학소설의 고전으로 손꼽힌다. 파리에서 출판된 이 소설은 무엇일까?

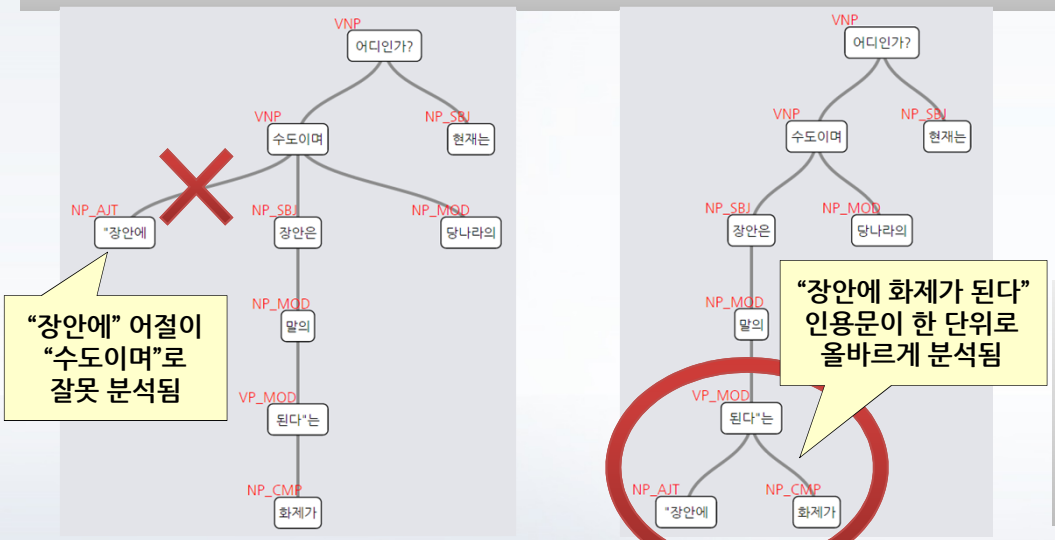


자연어의 특성을 고려한 실질적 기술 개발

언어의 생명성과 문장 부호의 의미를 이해 가능한 기술 개발

- 국립국어원 “2014년 표준어 추가 시정안” 발표 (2014.12.15.)
  - 신규 표준어 및 신조어의 지속적 학습을 통한 성능 개선 (WP2 지속적 학습 기술 연계)
- 국립국어원 “한글 맞춤법 문장 부호 개정안” 발표 (2014.10.27. 발표, 26년 만의 개정)
  - 인터넷 환경의 글쓰기에 적합하도록 문장 부호 용법 현실화: 조항 수 66개 → 94개 증가
  - 문장 부호 사용 빈도: 문장 당 평균 1.2회 발생 (위키백과)
  - 문장 부호로 표현되는 문장 구조 반영을 통한 의존구문분석 성능 개선: 90.1% → 91.2%

예문: "장안에 화제가 된다"는 말의 장안은 당나라의 수도이며 현재는 어디인가?



윌리엄 셰익스피어

위키백과, 우리 모두의 백과사전.

윌리엄 셰익스피어(William Shakespeare, 1564년 4월 26일[1]~1616년 4월 23일)는 영국의 극작가, 시인이다. 그의 작품은 영어로 된 작품 중 최고라고 찬사받고 셰익스피어 자신도 최고 극작가로 손꼽힌다.[2] 그는 자주 영국의 "국민 시인"과 "에이번의 시인"으로 불렸다.[3]

윌리엄 셰익스피어(William Shakespeare,
1564년 4월 26일 ~1616년 4월 23일)는 영국의 극작가, 시인이다.

→ 자연어 이해 대상 문장: 윌리엄 셰익스피어는 영국의 극작가, 시인이다.
→ 문장 부호 기반 구문 처리
William Shakespeare (type: 부가설명)
1564년 4월 26일 ~1616년 4월 23일 (type: 연대)

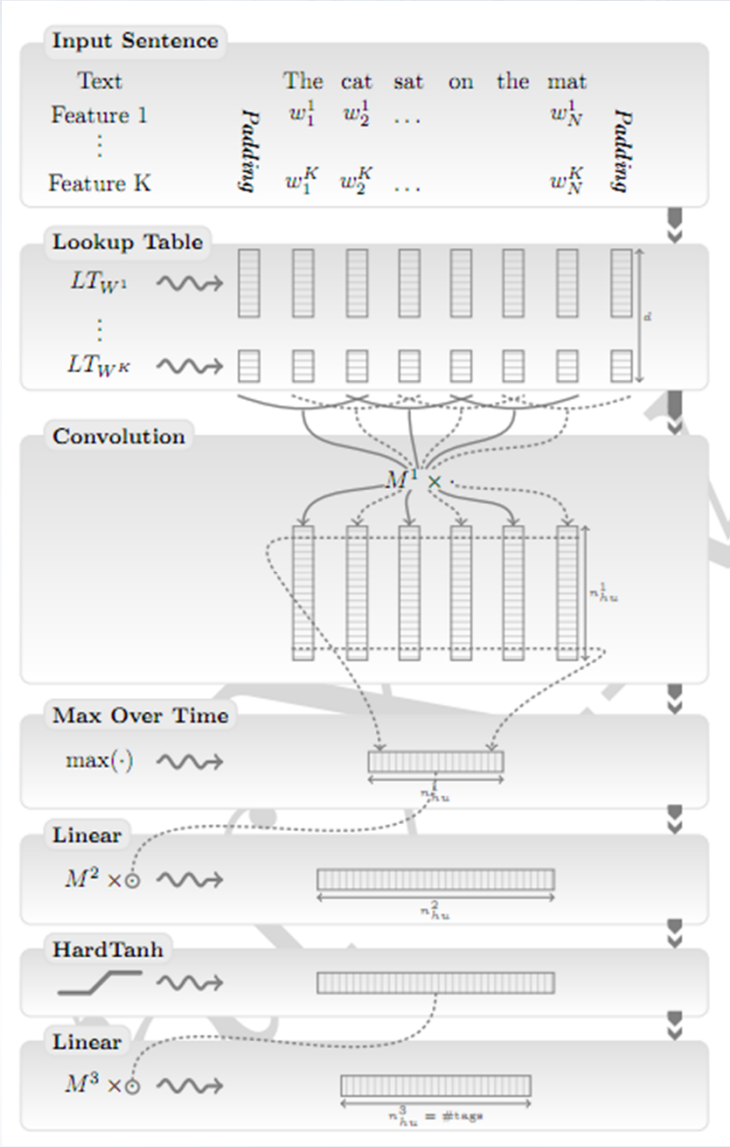
〈문장부호 반영 이전 구문분석 결과〉 〈문장부호 반영 이후 구문분석 결과〉

Deep Learning 적용 자연어 심층이해 성능 개선

- **접근방법**
  1. Deep Neural Network(DNN)
     - 입력으로 Pre-training된 결과 사용
       - 예: Word Embedding, Phrase Embedding
     - 구조적 분류를 위한 신경망 구조 확장 적용
       - Convolutional Neural Network, LSTM
  2. 지도학습 방법에 WE 결과를 학습 자질로 사용
     - Word2Vector 이용하여 학습
     - K-means 이용하여 클러스터링

- **실험결과**

방법론	개체명 인식	의존 구문분석	의미역 인식	상호참조 해결
지도 학습	90.7%	92.5%	77.8%	60.46%
DNN	88.4%	90.4%	75.1%	69.6% (No-pretraining: 65.8%)
지도 학습 + WE	89.0%	-	76.9%	-



〈Sentence approach network, R. Collobert/JMLR 2011〉

LSTM:Long Short-Term Memory

- RNN architecture specifically designed to address the vanishing gradient and exploding gradient problem
- The hidden neural units are replaced by a number of memory blocks.
- Each memory block contains several cells whose activations are controlled by 3 multiplicative gates
- Input/forget/output gate

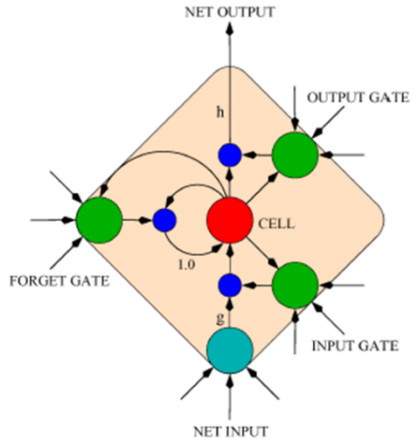


Figure 1: LSTM memory block with a single cell. (Graves et al., 2009)

- Deep Bidirectional LSTM
  - A standard LSTM processes the sequence in forward direction.
  - The output of this LSTM layer is taken by the next LSTM layer as input, processed in reversed direction
  - These two standard LSTM layers compose a pair of LSTM

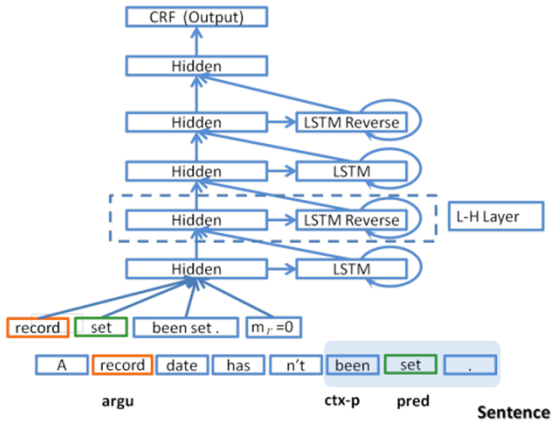


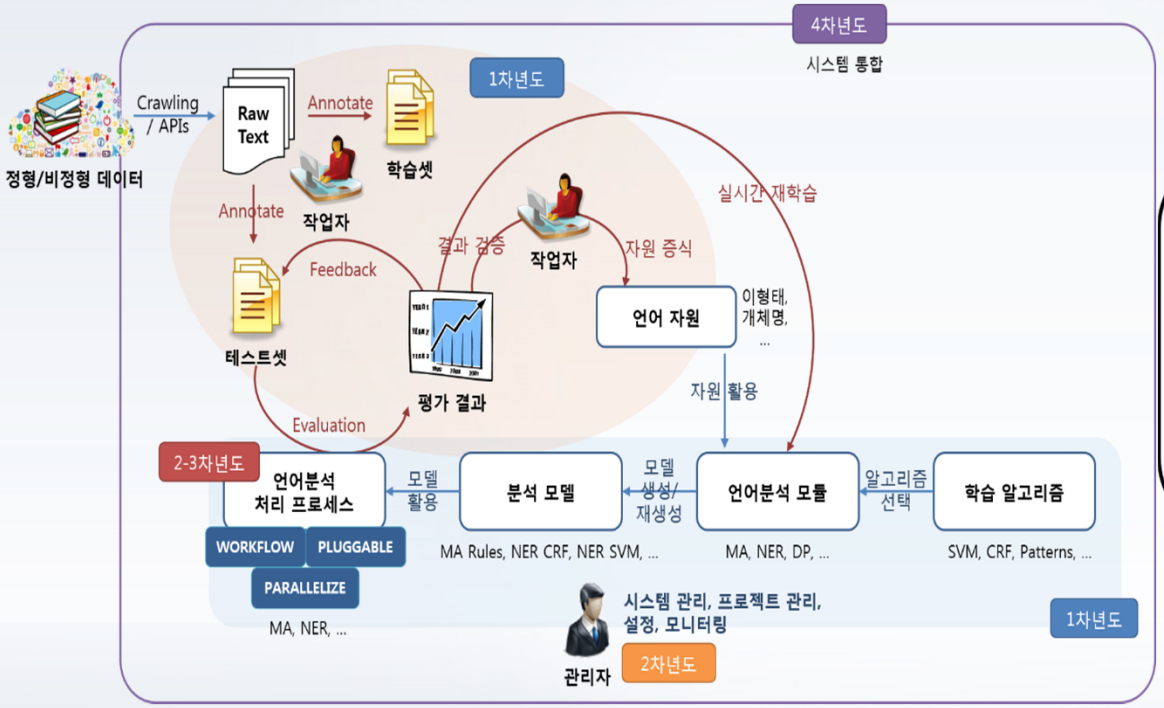
Figure 2: DB-LSTM network.Shadow part denote the predicate context within length 1.

End-to-end Learning of SRL Using RNN (Zhou, 2015)

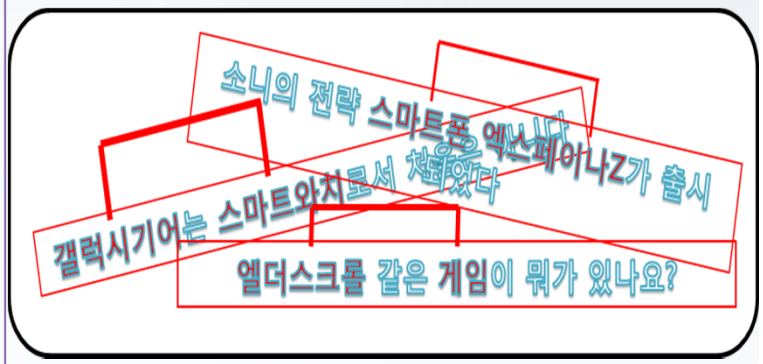
지속적 학습 목표 및 구성도

학습지능: 빅데이터로부터 끊임없이 언어지식 학습 및 증강

- 방법론: 빅데이터로부터 끊임없이 언어지식을 추출하고 학습하는 기술 설계 및 방법론 정립
  1. Never-ending Language Learning: 빅데이터로부터 끊임없이 언어지식을 자가학습
  2. Language sustainability: 새로운 언어지식 획득
  3. Domain adaptability: 도메인 확장



<지속적 언어지식 학습 프레임워크>



<어휘지식 학습 예>

지속적 학습 적용 예

단서 문장

… 그러나, <COUNTRY:일본>과 <COUNTRY:미국>의 <POLICY:가쓰라-태프트 밀약> 후 변절한 이 사람은 누구인가?

단서기반 언어지식 확장

clue word 학습

협약, 조약

협정, 약조, 선언

화약, 밀약

언어이해 패턴 학습

<COUNTRY>과/와 <COUNTRY>의 <TARGET:POLICY>

언어이해 기술 개선

…지난 42년간 미궁 속에 묻혔던 <COUNTRY:한국>과 <COUNTRY:일본>의 <POLICY:'독도밀약'>의 실체가 드러났다. 월간중앙은 …

…1670년에 <COUNTRY:잉글랜드>와 <COUNTRY:프랑스>의 <POLICY:도버밀약>에 따라 1672년 특별 사면권을 …

학습 기반 언어지식 확장

언어자원(사전) 증강

난징조약, 을사조약, 강화도조약, …
독도밀약,도버밀약, 하로밀약, 비외르쾨밀약,

통계 기반 학습데이터 추가 및 언어이해 모델 증강

.. <POSITION:박정희 대통령>이 <COUNTRY:일본>과 <POLICY:독도밀약>을 했는데, 아마도 <ISLAND:독도>를 …

<DATE:1965년 1월> <COUNTY:성북동>에서 <POSITION:정일권 국무총리>와 <COUNTRY:일본> <POLITICS:자민당>의 실력자 <POSITION:우노 소스케 의원>이 <POLICY:독도 밀약>에 사인을 한 사건이 있었습니다

학습 기반 언어이해 기술 개선

<POLICY:비외르쾨밀약>은 <DATE: 1905년 7월 24일> <POSITION: 러시아 황제 니콜라이 2세>와 <POSITION: 독일 황제 빌헬름 2세>가 맺은 비밀조약으로, …

<POLICY:도버밀약>은 <WAR:제3차 영국-네덜란드 전쟁> 당시 <COUNTRY:영국>과 <COUNTRY:프랑스>가 <DATE:1670년 6월 1일> 체결한 비밀 조약이다.

학습 기반 언어지식 확장

지식베이스 지식 증강

독도밀약: …
도버밀약: …
한로밀약: …
비외르쾨밀약: …
.
.

언어이해 패턴 학습

<COUNTRY>과/와 <COUNTRY>의 <TARGET:POLICY>

<TARGET:POLICY>은/는 <DATE> <POSITION>과/와 <POSITION>가 <맺, 체결하> 비밀조약

<TARGET:POLICY>은/는 <WAR> <COUNTRY>과/와 <COUNTRY>가 <DATE> <맺, 체결하> 비밀조약

…

어휘 의미관계(신조어,IS-A) 추출 기술 독창성

빅데이터로부터 언어지식을 끊임없이 학습하는 방법론 정립

* Continuous Learning 사이클 생성: "언어 지식의 자동 확장" → "언어 이해 성능 개선"

형태소 신조어

IS-A 관계

문서 수집

독일을 선택한 메수트 외질을 "국가의 배반자"라고 비판해, 한편 외질은 최근 4번의 캄프 누 원정에서 … 외질은 최근 EPL리그 뿐만 아니라 챔피언스리그에서도…

17,251개 어절

.. 외질(터키어: Mesut Ozil, 1988년 10월 15일 ~)은 잉글랜드 프리미어리그 아스널 소속의 축구 선수이다. … 외질은 분명 특별한 축구 선수로 평가된다. …

<시드 패턴 지식>

#INSTANCE#=NP[PS_NAME] → #TYPE#=VNP[CV_OCCUPATION](이다)

303,591개 위키백과 문서

지식 추출

격조사 기반 단어 추출

독일을 선택한 메수트 외질을 "국가의 배반자"라고 비판해, 한편 외질은 최근 4번의 캄프 누 원정에서 … 외질은 최근 EPL리그 뿐만 아니라 챔피언스리그에서도…

독일 1, 외질 3, 국가 1, 4번 1, 원정 1, 챔피언스리그 1

신조어 후보 1,555개 추출

반복적 IS-A 어휘 의미관계 지식 확장

외질은 잉글랜드 프리미어리그 아스널 소속의 축구 선수이다. … → 통계 기반 신뢰도 필터링 → … 외질은 분명 특별한 축구 선수로 평가된다. 그가 제 기량을 발휘한다면, 아스널로선 천군만마를 얻은 것이나 … → 통계 기반 신뢰도 필터링

<확장 지식>

#INSTANCE#=NP[PS_NAME] → #TYPE#=NP_AJT[CV_OCCUPATION](로) → VP(평가되다)

IS-A 지식 551개 구축

지식 검증

사전 기반 필터링

독일 1, 외질 3, 국가 1, 4번 1, 원정 1, 챔피언스리그 1

형태소 신조어 사전 생성

외질 NNG, 챔피언스리그 NNG

신조어 176개 추출

자동평가 기반 신조어 검증

신조어 추출정확도 94.7%

성능하락: 아니오 / 성능하락: 예

외질 NNG 챔피언스리그 NNG

외질 NNG 챔피언스리그 NNG

고빈도 저신뢰 지식 필터링

#INSTANCE#=NP[PS_NAME] → #TYPE#=NP_AJT[CV_OCCUPATION](로) → VP(평가되다)

추출 어휘 의미관계 → (홍진영, 복면가왕), (공지영, 노동운동가)

IS-A 303,268개 추출

자동평가 기반 IS-A 어휘 의미관계 지식 검증

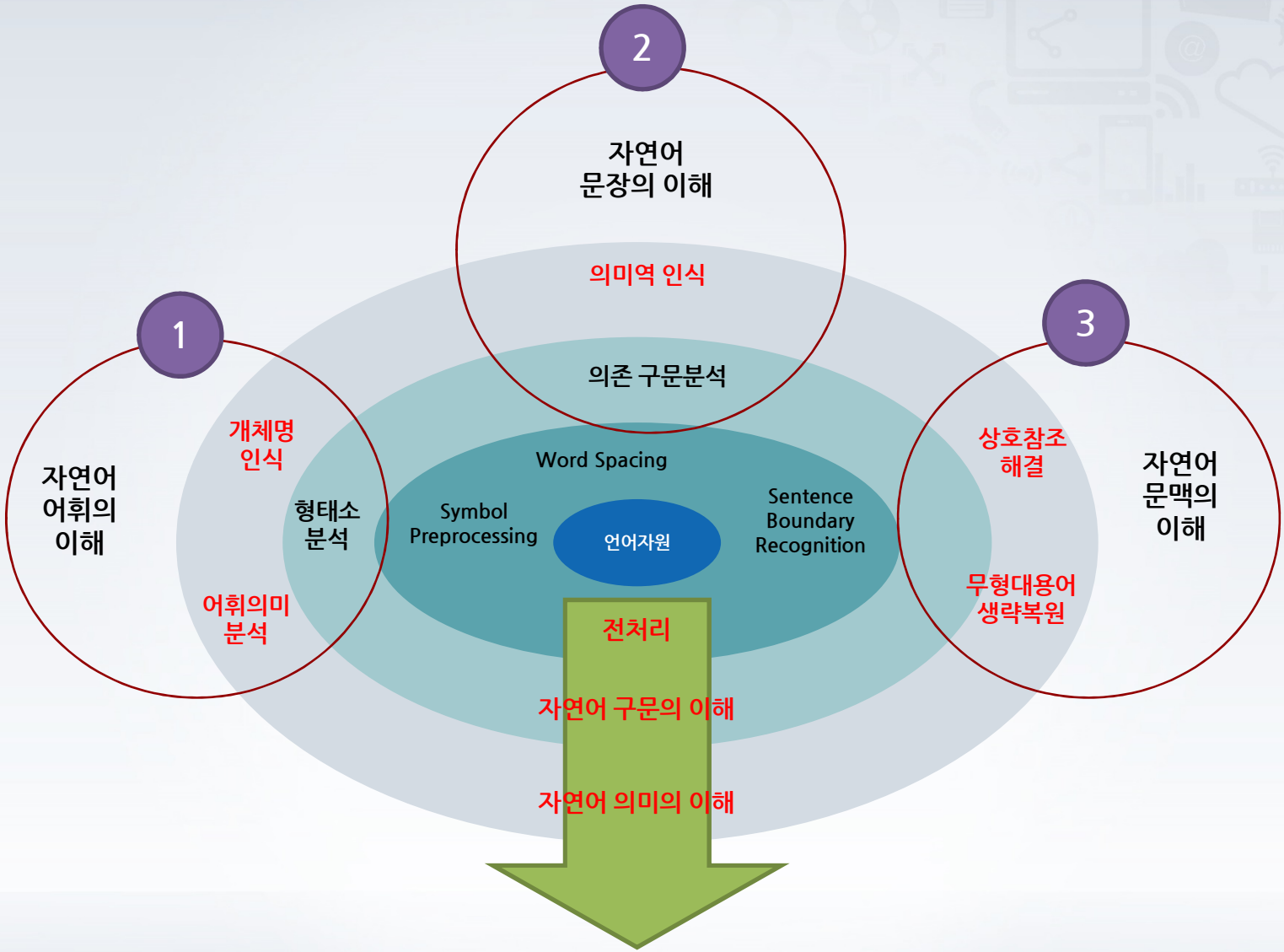
IS-A 추출정확도 89.3%

성능향상: 예 / 성능향상: 아니오

#INSTANCE#=NP[PS_NAME] → #TYPE#=NP_AJT[CV_OCCUPATION](로) → VP_[평가되다]

#INSTANCE#=NP[PS_NAME] → #TYPE#=NP_AJT[CV_OCCUPATION](로) → VP_[평가되다]

WiseNLU: 개발방향



Broad-Coverage Semantic DP (SemEval 2014 Task8)

- Syntactic/Semantic representation 비교

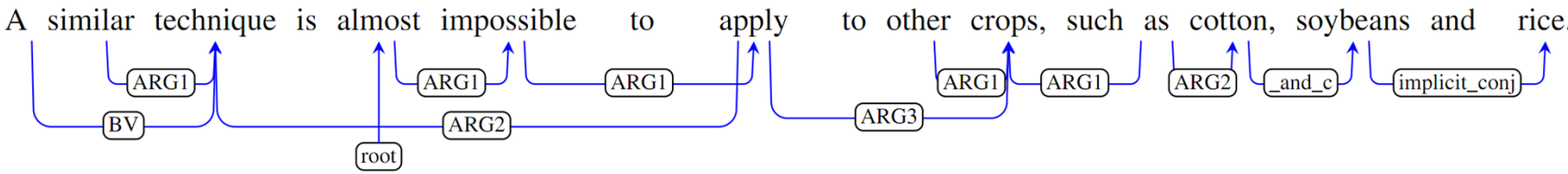
	Syntactic DP	Semantic DP(SRL)
접근성	Root node로부터 어떤 node든지 접근 가능	유일한 root가 존재하지 않으며, 접근 불가능한 node도 존재
Path 유일성	Root node로부터 특정 node까지 유일한 path만 존재	특정 node에 접근할 수 있는 path가 여러 개 존재할 수 있음

- Semantic representation을 위해서 general graph processing을 도입하려 함
  - 'who did what to whom'을 좀더 direct하게 표현 가능하도록
  - 의존 문법의 projectivity를 무시함(non-projectivity)
  - 궁극적으로 모든 content words을 1개의 구조로 통합하려 함 → 현재는 predicate 단위로 분리되어 있음

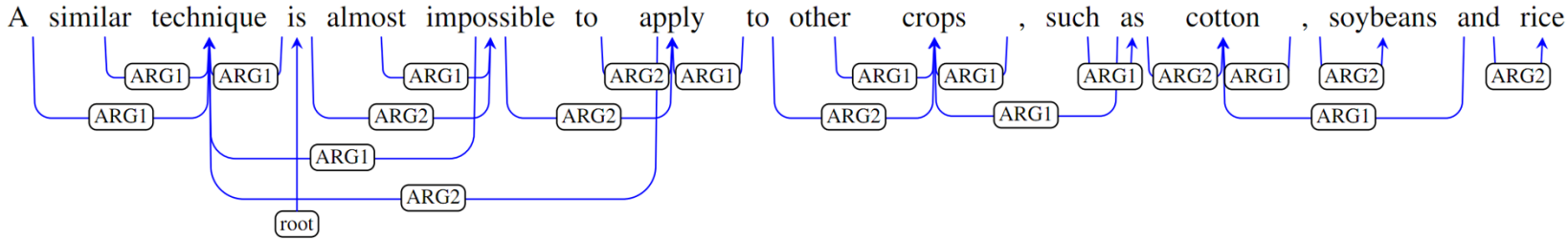
- 기존 PropBank NomBank는 verbal/nominal predicate에 대해 argument identification으로 국한됨
  - Negation, Scopal embedding, Comparatives, Possessives, Various types of modification, conjunction

Semantic DP Representations

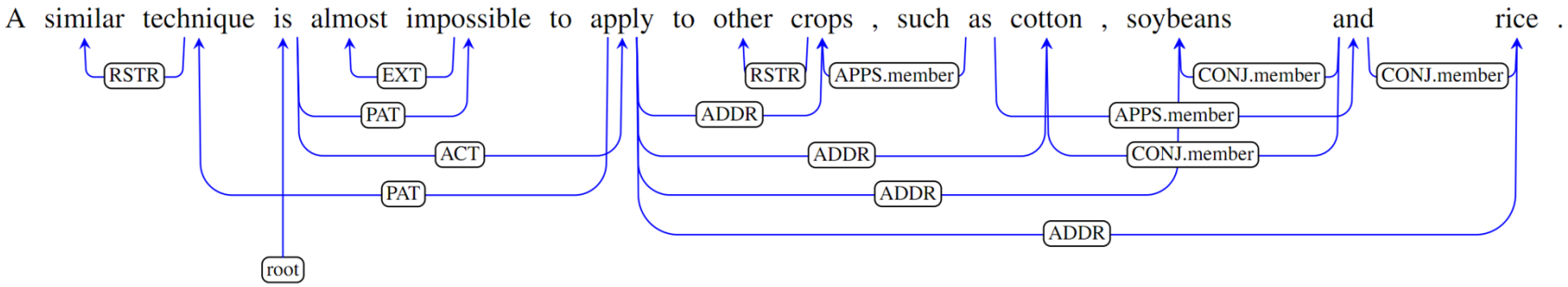
Example 1: CoNLL 2008 semantic roles (from PropBank and NomBank).



Format 1: Minimal Recursion Semantics–derived dependencies (DM), from DeepBank HPSG annotations.



Format 2: Predicate-argument structures (PAS), from Enju HPSG annotation.



Format 3: Parts of the tectogrammatical layer of the Prague Czech-English Dependency Treebank (PCEDT).

Abstract Meaning Representation(Banarescu et al, 2013)

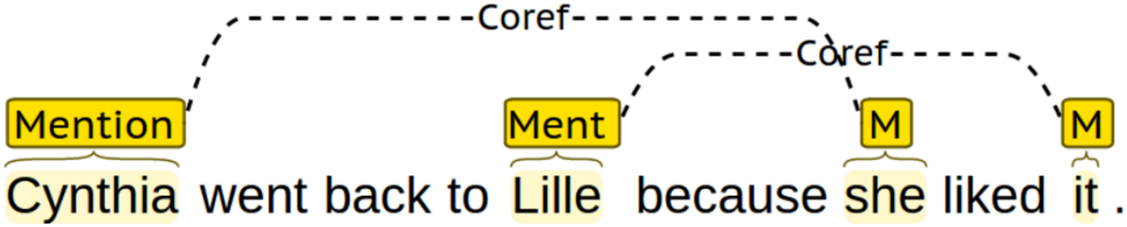
- Motivation: unify all semantic annotation

Semantic role labeling:

- predicate + semantic roles

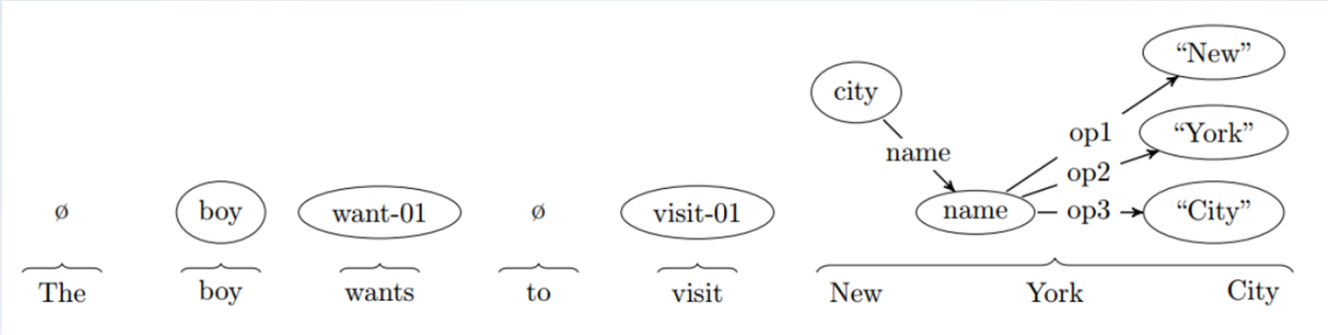
Named-entity recognition:

Coreference resolution:

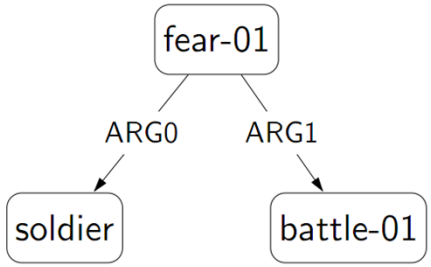


Abstract Meaning Representation(Banarescu et al, 2013)

- Sentence-level annotation



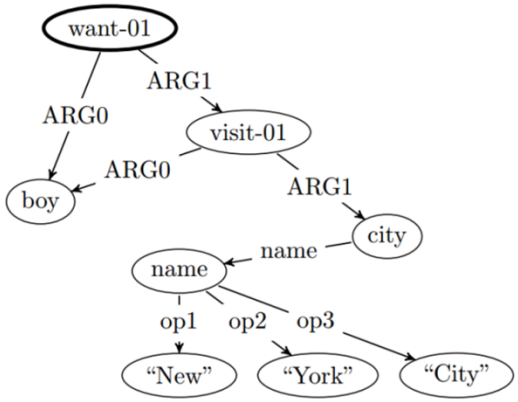
The soldier feared battle.
The soldier was afraid of battle.
The soldier had a fear of battle.
Battle was feared by the soldier.
Battle was what the soldier was afraid of.



Input: sentence

The boy wants to go to New York City.

Output: graph



WiseNLU: 국내 표준화 계획

- 한국어 이해 표준화 계획
  - 2차년도 표준화 결과
    - 대상기술: 형태소 품사세트
    - 표준 심의 및 공고 완료
  - 3차년도 표준화 계획
    - 대상기술: 개체명인식, 의존구문분석
    - 일정:

  - 4차년도 표준화 계획
    - 목표기술: 어휘의미분석, 의미역인식, 상호참조 및 무형대용어 복원

WiseNLU: 말뭉치 배포 계획

- WiseNLU 말뭉치 배포 계획
  - 개체명 태깅 말뭉치 배포 계획
    - 목표: 기계학습을 이용하여 학습이 가능한 수준의 말뭉치 공개
      - 2015년 5,000문장, 2016년 5,000문장 배포 목표
    - 태그셋: PLO + Misc. 또는 ETRI 태그셋 대분류 사용
    - 배포시기: 한글및한국어정보처리학회 (10월16일~17일)
    - 비고: 차년도 국어 정보 처리 시스템 경진 대회에 ETRI 말뭉치 활용 예정
  - WiseQA 평가셋 배포 계획
    - 대상 코퍼스: 형태소분석, 개체명인식, 어휘의미분석, 구문분석, 의미역 인식
    - 배포시기: 한글및한국어정보처리학회 (10월16일~17일)
    - 배포대상 코퍼스: GS3.0
    - 소스 콘텐츠: 장학퀴즈 질문 및 정답 단락 (위키백과, 표준국어대사전, 등)
    - 배포 수량
      - 약 500~600문장 수준으로 예상
    - 태깅 가이드 매뉴얼
      - 언어분석 표준화 제안과 동일한 가이드라인 적용

WiseNLU 배포 현황 및 계획

- WiseNLU 자연어이해기술 배포
  - 대상기술: 7개 기술
    - 형태소 분석 기술, 어휘의미분석 기술, 개체명인식 기술, 구문분석 기술, 의미역인식 기술, 상호참조해결 기술, 무형대용어 복원 기술
  - 배포기관: 대학 14개 연구실, 기업 4개
  - 3차년도 배포 및 계획
    - 3월 초: 3차년도 1차 시스템 제공 (배포완료)
    - 6월 중순: 신규 모듈 추가 및 자료구조, 활용편의성 개선 (배포완료)
    - 9월 중순: 주요 기술 별 성능 개선 버전 배포
    - 12월 초: 신뢰성 개선 및 3차년도 최종 WiseNLU 시스템 배포

감사합니다.