

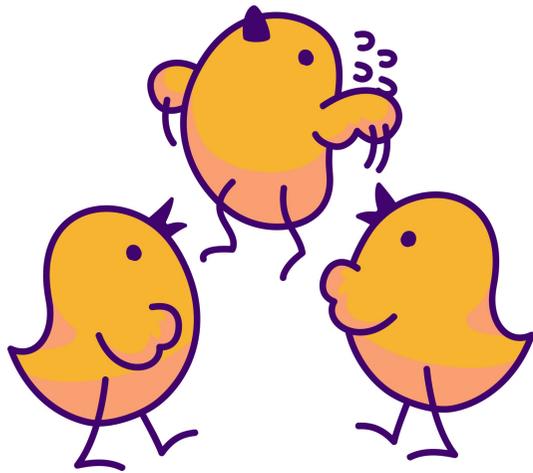


LangCon 2020



Spoken Language Understanding의 이해와 사례: End-to-End SLU

NCSOFT Speech AI Lab 음성인식팀 송치성

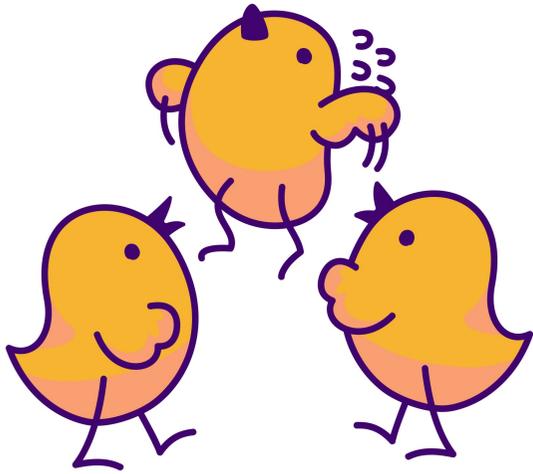


- **Voice User Interface**
- **Spoken Language Understanding**
- **Process of SLU**
- **Case Study : End-to-End SLU**



○ 이 발표에서 다루지 않는 것

- 오디오 신호처리
- 음성 피쳐 추출 방법
- 음성인식 프로세스
- 복잡한 수식
- 세세한 코드 리뷰



Voice User Interface

LangCon 2020

Voice User Interface

“미래의 디지털 시스템 인터페이스는 더 이상 기계에 의한 것이 아닐 것이다. 인간 중심적(Human-centric)일 것이다.”

“우리는 디지털 시스템을 위한 자연스러운 인터페이스(NUI)를 구축할 수 있고, 그것으로 모든 환경이 활성화 될 것이다.”

- 아마존 CTO, Werner Vogels

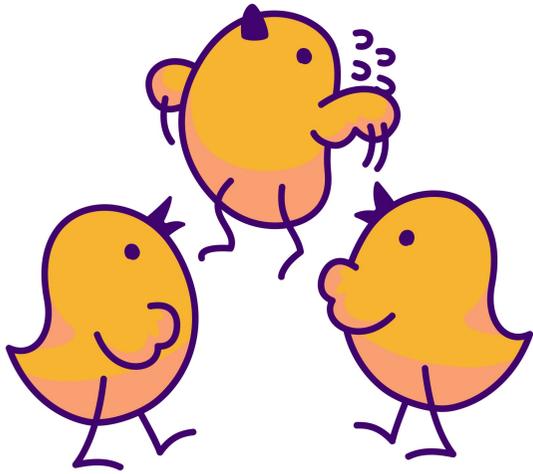




○ Voice User Interface (VUI)

- 컴퓨터 시스템과 사용자의 음성을 기반으로 하는 인터페이스
- 인간 중심적 (Human-centric)이고 자연스러운 인터페이스(NUI; Natural User Interface):
음성은 인간에게 가장 기본적인 커뮤니케이션 수단이며 보편적인 인터페이스이며, 기존의 인터페이스(CLI, GUI 등)에 비해 자연스러움
- 스마트폰, 스마트스피커, 스마트TV, 자동차 네비게이션 등에 접목되어 활용

- Hands-free : 인터랙션을 하는 동안 주의를 기울이지 않아도 되고, 양손이 자유로움 (멀티태스킹 가능)
- Zero-depth : GUI는 계층적이어서 일련의 액션을 통해 목적을 이루지만, VUI를 통해서는 여러 단계를 거치지 않고 즉각 실행
- Fast : 타이핑보다 음성인식을 통한 입력이 빠름. (영어와 중국어의 경우, 약 3배의 속도)
- Short Learning Curve : 직관적이기 때문에 높은 접근성을 가지며, 디지털 기기에 대한 숙련도가 낮은 계층(중,노년층 등)도 쉽게 사용 가능
- WOW Factor : 기존에 없던 새로운 인터랙션을 제안하여 사용자를 감탄하게 하는 요소가 될 수 있음



Spoken Language Understanding

LangCon 2020

Spoken Language Understanding

Language
Conference





○ Spoken Language Understanding (SLU)

- 대화 시스템에서 발화로부터 의미(Meaning)나 의도(Intent)를 추론
- 추론된 결과는 미리 정의된 의미 표상(Semantic Representation) 포맷으로 표현하여 컴퓨터 시스템이 적절한 기능을 수행 할 수 있게 함
- 일반적으로 다음 두가지 태스크로 나눌 수 있음
 - Intent Detection : 발화자의 의도 파악
 - Slot Filling : 해당 의도를 해결하는데 필요한 Slot 파악

- 음성 언어 이해의 어려움

**Extra-
grammaticality**

Disfluency

**Speech
Recognition
Error**

**Out-of-domain
Utterance**

○ Extra-grammaticality

- 말을 할 때에는 글로 쓸 때와 다르게 잘 짜여진 형태(well-formed)가 아님.
- 일반적으로 문어에서보다 구어에서 문법적 제약을 덜 받음.



엄마 나 물!

○ Extra-grammaticality

- 말을 할 때에는 글로 쓸 때와 다르게 잘 짜여진 형태(well-formed)가 아님.
- 일반적으로 문어에서보다 구어에서 문법적 제약을 덜 받음.

○ Disfluency

- 일상에서의 대화는 유창하지 않고, 간투어(있지, 참, 자!), 말더듬, 머뭇거림, 반복 등이 포함되어 있음.



<https://www.youtube.com/watch?v=Fnnahpetxkc>

○ Extra-grammaticality

- 말을 할 때에는 글로 쓸 때와 다르게 잘 짜여진 형태(well-formed)가 아님.
- 일반적으로 문어에서보다 구어에서 문법적 제약을 덜 받음.

○ Disfluency

- 일상에서의 대화는 유창하지 않고, 간투어(있지, 참, 자!), 말더듬, 머뭇거림, 반복 등이 포함되어 있음.

○ Speech Recognition Error

- 일반적으로 SLU는 음성인식에 의존적
- 음성인식기술은 환경 잡음, 화자의 특성, 도메인 적응 기술 등에 영향을 받고, 이때의 오류가 전파됨

234174 아만자가 모에여?

Good Bad 0 (0명) gya|szzang

답변이 완료된 질문입니다. (2003-10-30 21:26 작성)

로즈마리라는 드라마에서 유희열이 "나 아만자여"라고 하는데 아만자가 모에여?

질문자가 선택한 답변

--

onewayand (2003-10-30 21:47 작성)

질문자 평 : ㅋㅋ 내 귀가 먹었나봐여 -- ★★★★★

암환자, 암!환자 암걸린사람입니다=_=



○ Extra-grammaticality

- 말을 할 때에는 글로 쓸 때와 다르게 잘 짜여진 형태(well-formed)가 아님.
- 일반적으로 문어에서보다 구어에서 문법적 제약을 덜 받음.

○ Disfluency

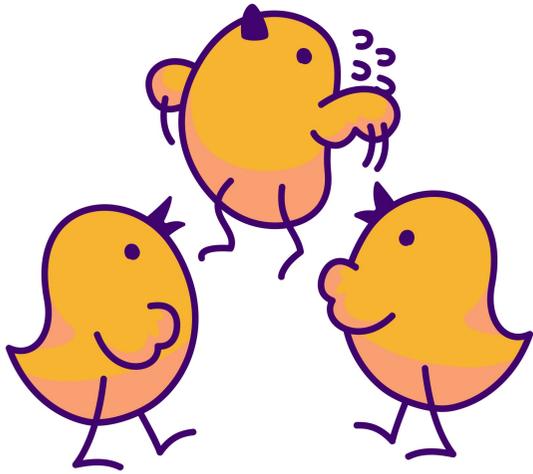
- 일상에서의 대화는 유창하지 않고, 간투어(있지, 참, 자!), 말더듬, 머뭇거림, 반복 등이 포함되어 있음.

○ Speech Recognition Error

- 일반적으로 SLU는 음성인식에 의존적
- 음성인식기술은 환경 잡음, 화자의 특성, 도메인 적응 기술 등에 영향을 받고, 이때의 오류가 전파됨

○ Out-of-domain Utterance

- SLU 시스템은 도메인에 한정적인데 반해 대화 시스템은 발화 내용의 제한을 두지 않음
- 도메인 밖의 발화에 대해 적절한 배제가 필요

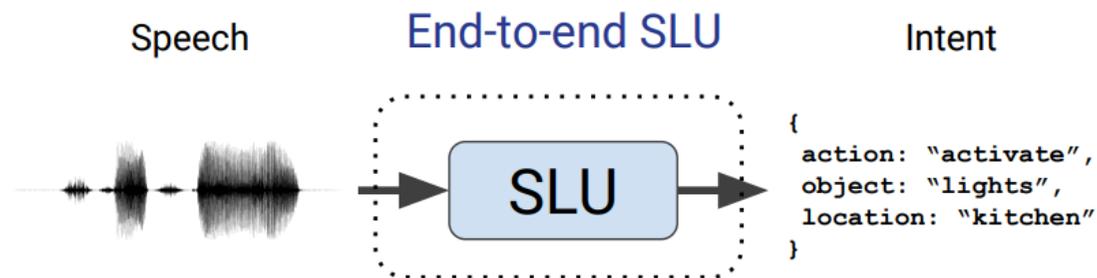
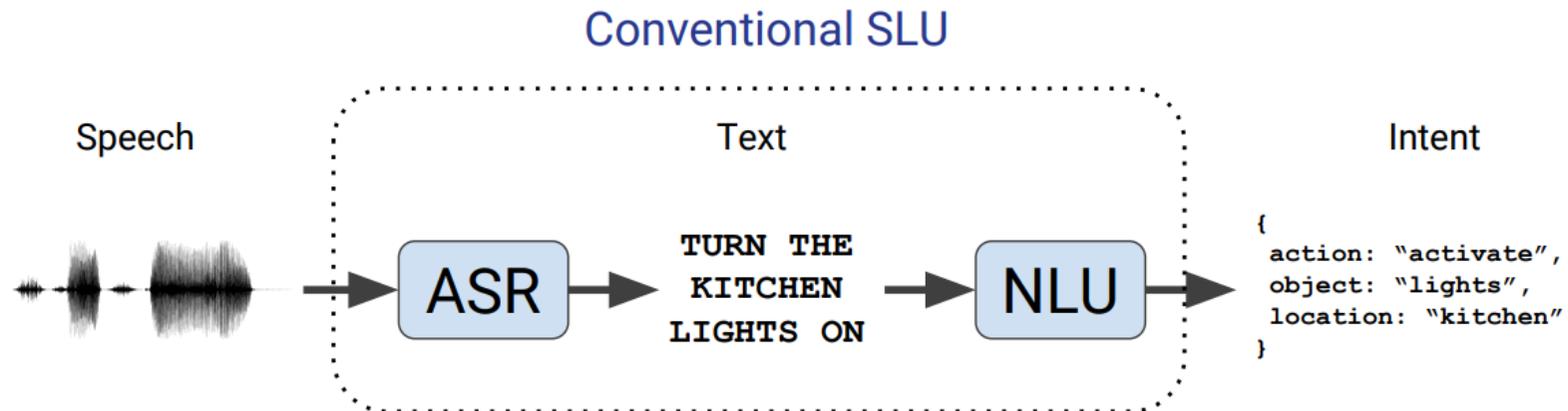


Process of SLU

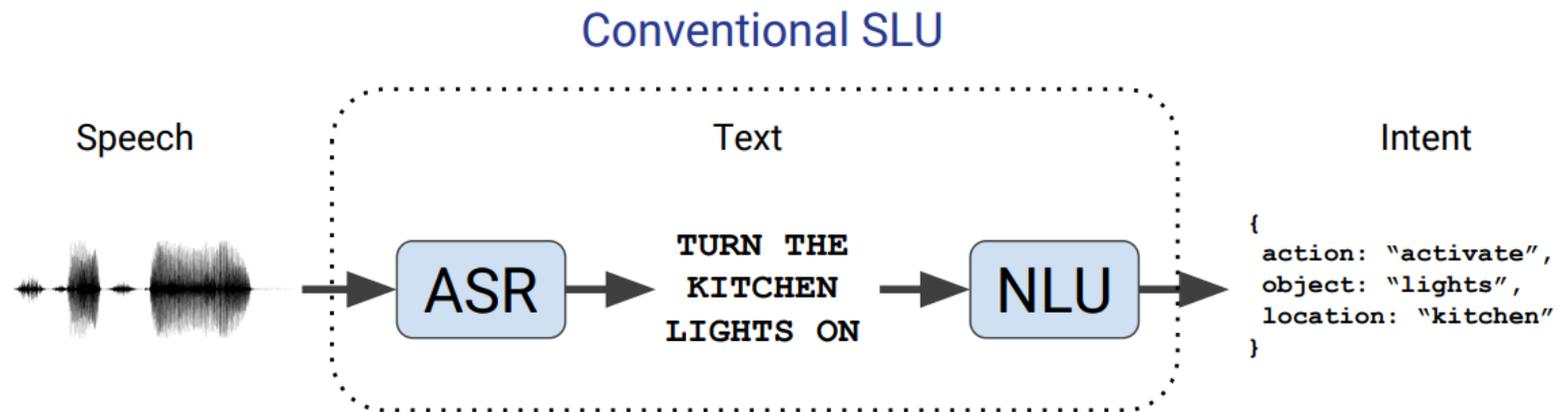
Process of SLU

○ 음성 언어 이해 프로세스

- 1. Conventional SLU
- 2. End-to-end SLU



○ 음성 언어 이해 프로세스 - 1. Conventional SLU



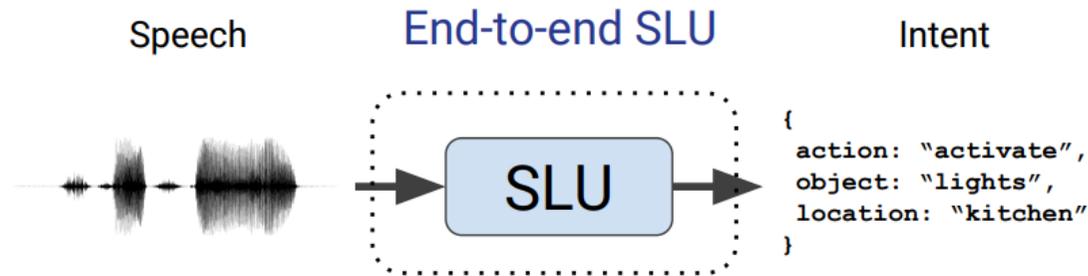
- 음성인식(ASR) 모듈과 자연어 이해(NLU) 모듈이 연결된 파이프라인 구조
- ASR 모듈을 통해 문자열로 전사된 발화에 대해 의미 표상(Semantic Representation) 추론
- 파이프라인 구조이기 때문에 음성인식기에 의존적 (음성인식 오류 전파)
- Rule-based / Data-driven 접근방법

○ 음성 언어 이해 프로세스 - 1. Conventional SLU

- Rule-based / Data-driven 접근방법
 - Rule-based Approach : 정규표현식 등의 간단한 규칙을 미리 정의하여 패턴 매칭
 - Data-driven Approach : Sequence labeling 문제의 관점에서 확률 모델 학습 (CRF 등 사용)

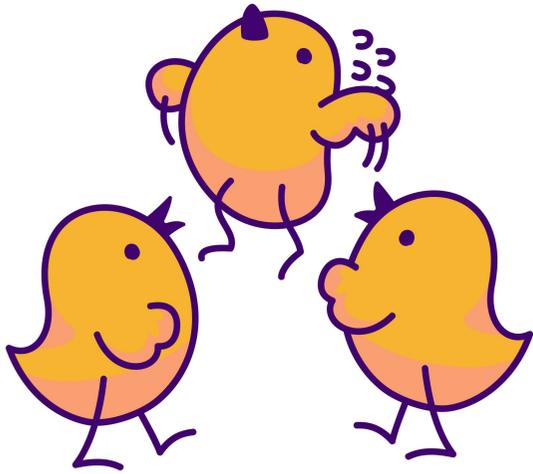
Utterance	show	flights	from	Seattle	to	San	Diego	tomorrow
Slots	O	O	O	B-fromloc	O	B-toloc	I-toloc	B-depart_date
Intent	Flight							

- 사용성 관점에서의 이슈:
 - 사용자는 높은 인식 정확도를 보장하기 위해 시스템이 학습한 정확한 단어 또는 구문을 말해야 함.
But, 그러기위해서 사용자가 시스템이 커버 가능한 명령을 외워야하기 때문에 불편
(사실 보통 이려고 있음..)
 - 이 문제를 극복하기 위해 시스템은 각 의도를 가진 여러 변형된 명령을 처리 할 수 있어야 함.
-> 사용자가 패러프레이징하여 유사한 대체 명령을 커버할 수 있음.



○ 음성 언어 이해 프로세스 - 2. End-to-end SLU

- 각각의 모듈을 각각의 목적으로 최적화 하는게 아니라 음성 피쳐부터 의미 표상까지 직접적으로 최적화
- 텍스트 형태로 전사하려는 불필요한 노력을 줄임으로써 간결한 모델을 만들 수 있고, 그 중간 과정(탐색알고리즘, 언어모델, FST 등..)에서 오류가 발생 할 수 있는 여지를 최소화
- 운율(prosody)과 같은 텍스트에는 없는 발화의 정보를 활용하는 것이 가능
 - "지금 학교가" -> 지금 학교에 가니? / 지금 학교니?
 - "잘한다" -> 잘한다^^ / 잘~ 한다 (Sarcasm)



Case Study : End-to-end SLU

LangCon 2020

Speech Model Pre-training for End-to-End Spoken Language Understanding

*Loren Lugosch¹, Mirco Ravanelli¹, Patrick Ignoto²,
Vikrant Singh Tomar², Yoshua Bengio^{1,3}*

¹Université de Montréal / Mila, ²Fluent.ai

³CIFAR Fellow

{lugoschl, mirco.ravanelli, yoshua.bengio}@mila.quebec
{patrick.ignoto, vikrant}@fluent.ai

○ Speech Model Pre-training for End-to-End Spoken Language Understanding

- Interspeech 2019에서 발표
- Paper : <https://arxiv.org/abs/1904.03670>
- Code : <https://github.com/lorenlugosch/end-to-end-SLU>



Split	# of speakers	# of utterances	# hours
Train	77	23,132	14.7
Valid	10	3,118	1.9
Test	10	3,793	2.4
Total	97	30,043	19.0

○ 'Fluent Speech Commands' Dataset

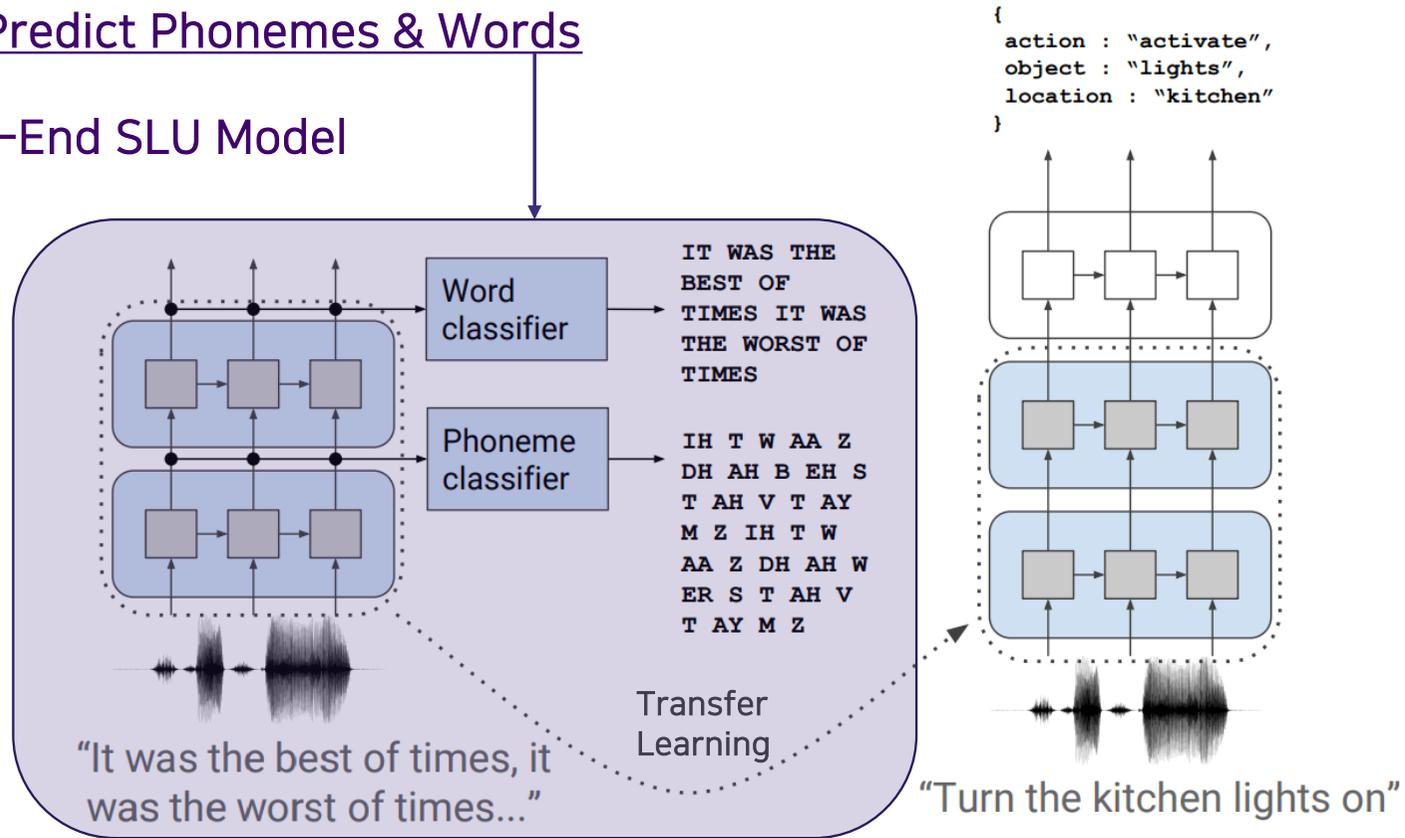
<데이터셋 규모>

- Fluent.ai에서 학술 연구 목적으로 공개
- 총 97명의 화자, 약 3만 발화, 약 19시간의 규모
- 스마트홈 시나리오에서 Voice Assistant와 인터랙션하는 음성 명령 데이터셋
 - Ex) "Turn on the lights"
- Intent 구조 : (action, object, location) 3개의 벡터로 이루어진 간단한 의미 표상 (총 31개)
 - Ex) (activate, lights, none)
- 일반화를 위한 다양한 패러프레이징
 - "Switch on the lights", "lights on", ...
- Dataset : <https://fluent.ai/fluent-speech-commands-a-dataset-for-spoken-language-understanding-research/>

Procedure & Architecture

i) Pre-train to Predict Phonemes & Words

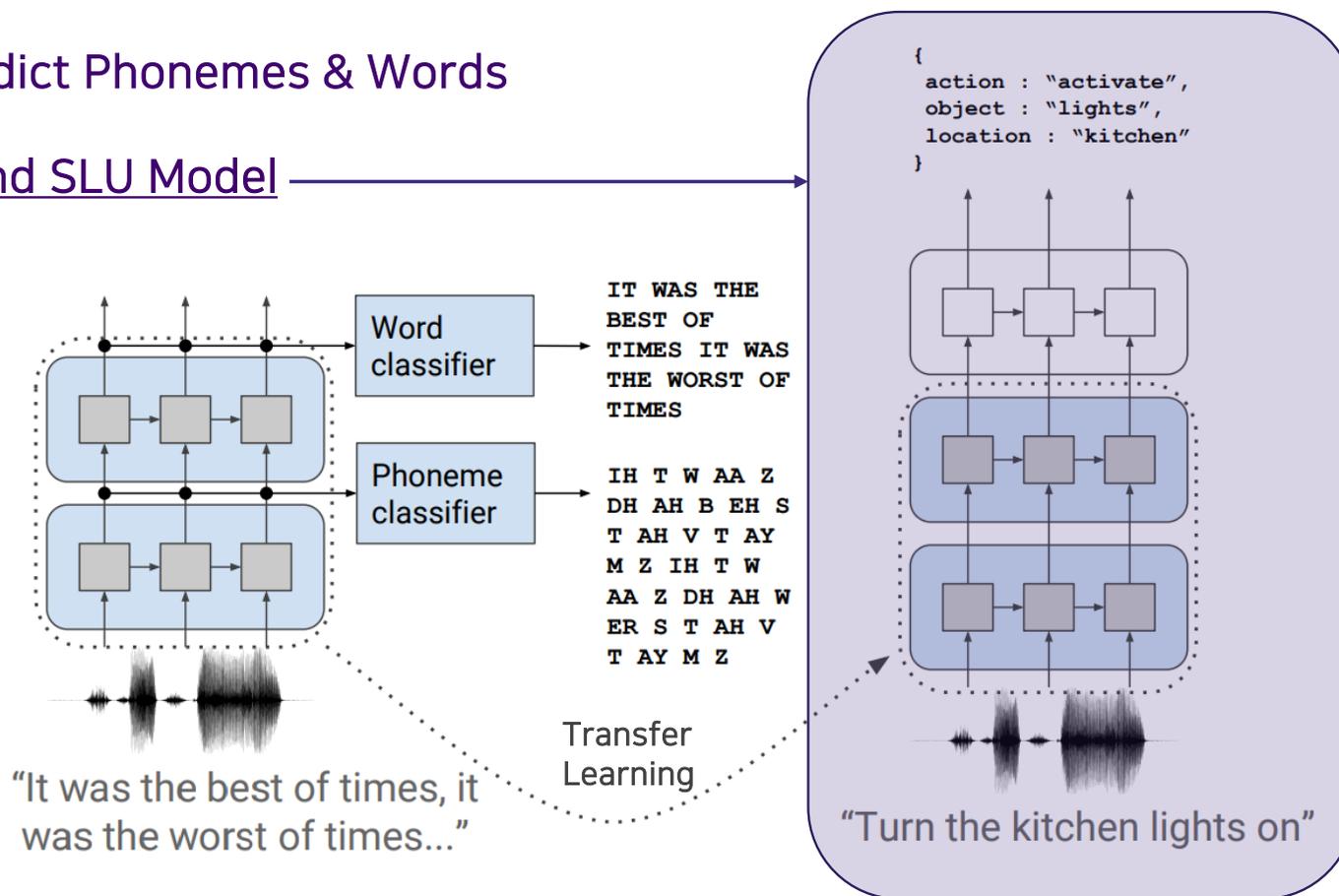
ii) Train End-to-End SLU Model

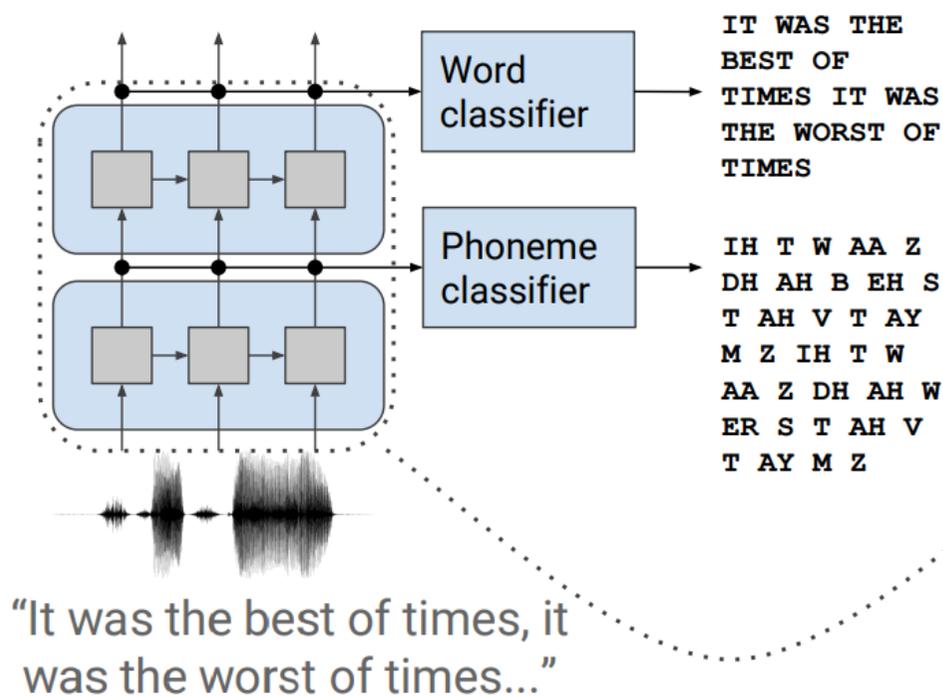


Procedure & Architecture

i) Pre-train to Predict Phonemes & Words

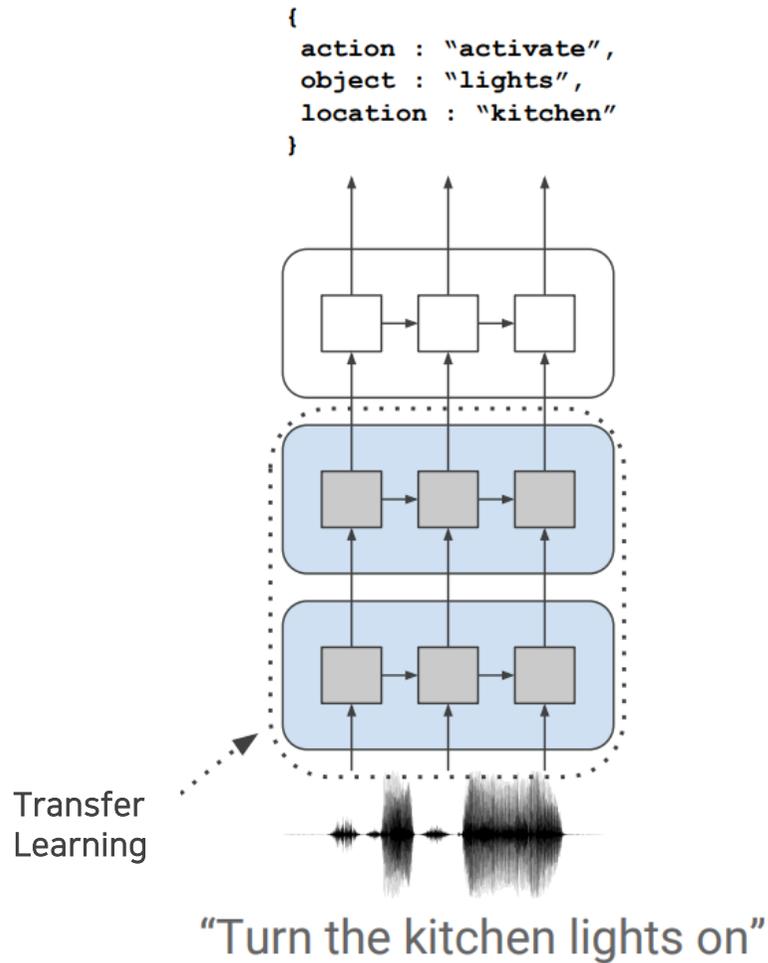
ii) Train End-to-End SLU Model





i) Pre-train to predict phonemes & words

- 무슨 말을 했는지 알아 듣는 능력(음성인식)을 학습
- 학습 데이터 : Librispeech 음성 데이터
(<http://www.openslr.org/12>)
- 구조 : Convolution 레이어와 recurrent 레이어를 반복해서 쌓아서 각 음소(phoneme)와 단어(word) 예측
- 본 논문의 github 리포에서 학습된 파라미터 제공
- 입력 파형에 대한 피쳐 추출 방법으로 NN기반 피쳐 추출 방식 중 한가지인 SincNet layer를 사용
- SincNet에 대한 자세한 설명 :
<https://ratsgo.github.io/speechbook/docs/neuralfe/sincnet>

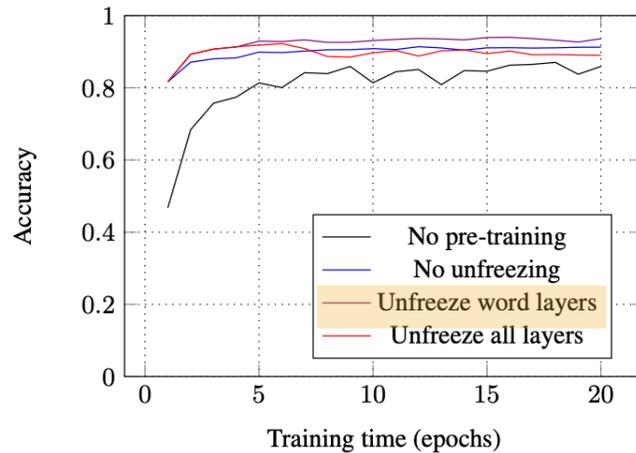


ii) Train End-to-End SLU Model

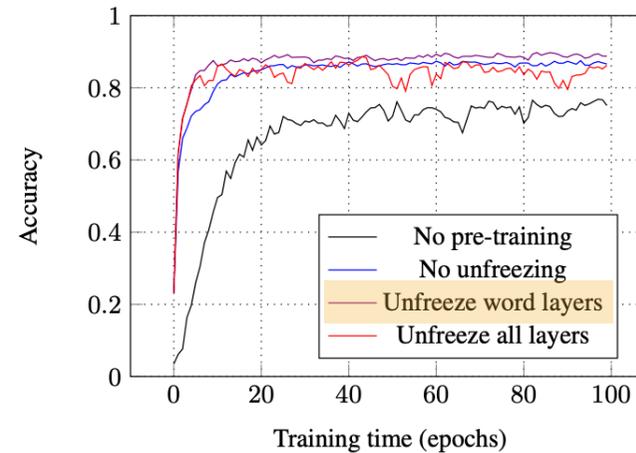
- 알아 들은 말이 무슨 의미인지 이해하는 능력을 학습
- 학습데이터 : Fluent Speech Commands
- 구조 :
 - 미리 학습한 음성인식 모듈에 intent 모듈을 결합
 - (action, object, location) 벡터로 이루어진 multi-label 형식의 의미 표상을 flatten하여 31개의 single-label 분류 방식으로 학습 (각 action, object, location은 독립적이지 않으며, autoregressive 하기 때문)
- Word/Phoneme 분류기 softmax 앞의 bottleneck 피쳐 사용
 - 1. 차원 수 증가 : 다음 모듈에 더 많은 정보를 가지고 감
 - 2. 컴퓨팅 리소스 감소

Experiments

- 전이학습을 통해 SLU 모델을 학습 할 때, word 레이어만 학습 가능하도록 한 것이 가장 성능이 좋음.



(a) Full dataset.



(b) 10% of the dataset.

Model	Accuracy (full)	Accuracy (10%)
No pre-training	96.6%	88.9%
No unfreezing	98.8%	97.9%
Unfreeze word layers	98.7%	97.9%
Unfreeze all layers	97.2%	95.8%



- Code : <https://github.com/daydrill/end-to-end-SLU-langcon2020>

- Lugosch, Loren, et al. "Speech model pre-training for end-to-end spoken language understanding." Interspeech, 2019.
- Tur, Gokhan, and Renato De Mori. Spoken language understanding: Systems for extracting semantic information from speech. John Wiley & Sons, 2011.
- <http://readmultiplex.com/2017/03/27/stanford-university-voice-first-is-3x-faster-than-typing/>
- <https://blog.naver.com/vinylx/221299546881>
- <https://ratsgo.github.io/speechbook/docs/neuralfe/sincnet>

Thank you - !

daydrilling@gmail.com | 송치성