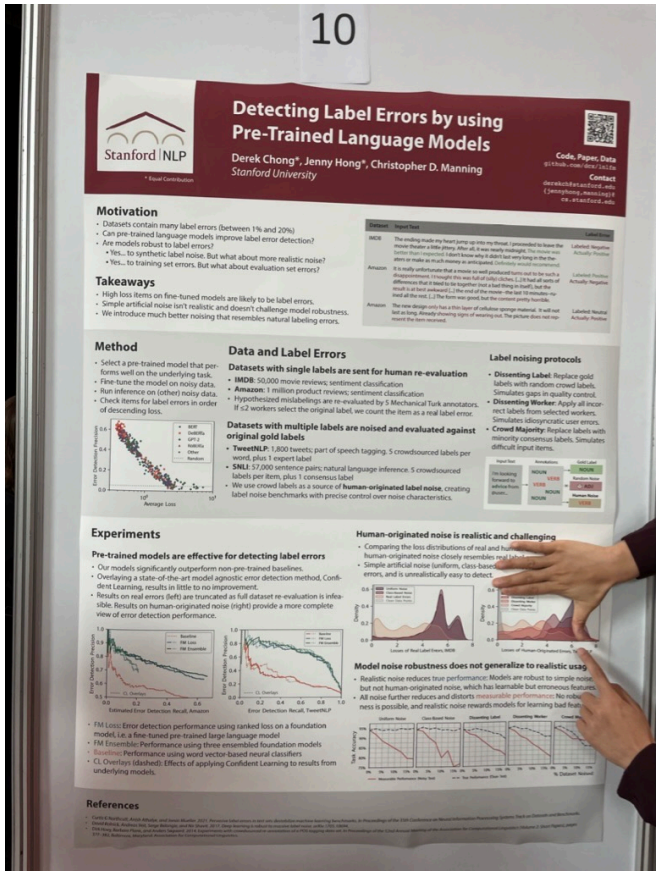
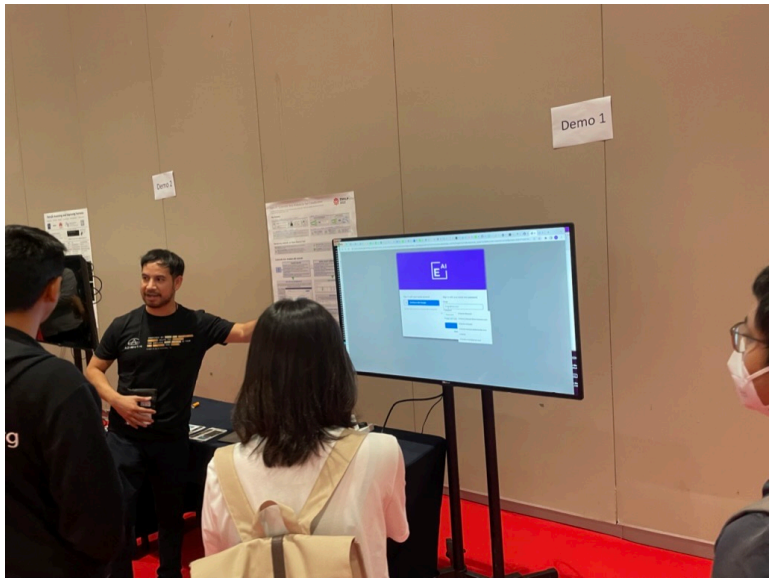
EMNLP 2022 세미나

발표자: 이승준

EMNLP 2022

- Synthetic Data & Error
- Empirical Experiments
- LMs is still important..?





LM-debugger 시현

Detecting Label Errors by using Pre-Trained Language Models

Derek Chong*
Stanford University
derekch@stanford.edu

Jenny Hong*
Stanford University
jennyhong@cs.stanford.edu

Christopher D. Manning
Stanford University
manning@cs.stanford.edu

Problem Statement

- 거의 모든 데이터셋에는 일반적으로 Label Noise가 존재
- 학습 관점에서 PLM는 특히 Robust하기 때문에 크게 문제가 되지 않음
- 하지만, Evaluation 단계에선 이것은 다른 문제
- Label Noise => Synthetic Noise: Uniform & Class-based

Introduction

- 거의 모든 데이터셋에는 일반적으로 Label Noise가 존재
- 학습 관점에서 PLM는 특히 Robust하기 때문에 크게 문제가 되지 않음
- 하지만, Evaluation 단계에선 이것은 다른 문제
- Label Noise => Synthetic Noise: Uniform & Class-based
- Contribution
  1. Pre-trained model can find REAL label error
  2. Human-originated noise: a more realistic and challenging benchmark, resembling nature labeling errors
  3. Evaluating with label errors

Background: Synthetic Noise

- **Class-based noise**: the errors are generated based on specific classes of errors that are commonly seen in real-world text data → randomly

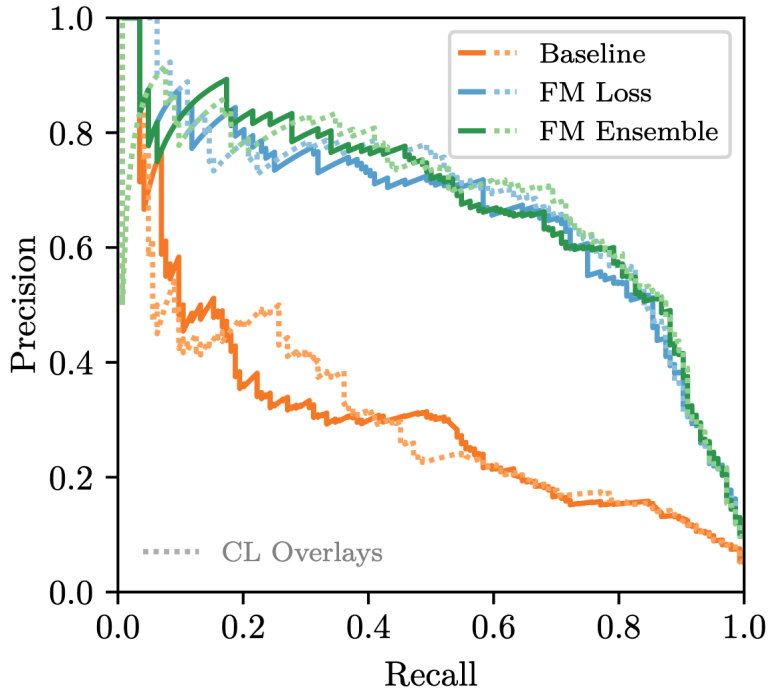
  Original text: "This movie is fantastic." with label "positive"
  Noisy text with label: "This movie is fantastic." with label "negative"

- **Uniform noise of label errors:** In this type of noise, the errors are generated uniformly across the labels, regardless of the type of error. → ex) uniform distribution

  Original text with label: "This movie is fantastic." with label "positive"
  Noisy text with label: "This movie is fantastic." with label "neutral"

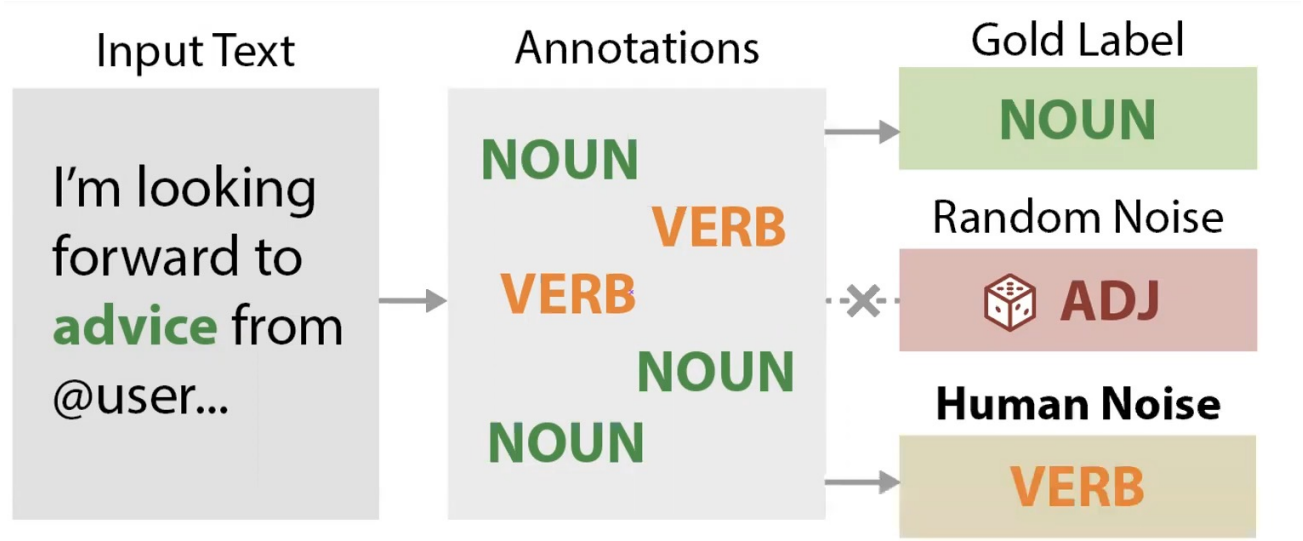
Pre-trained model find label errors

- PLMs may be a powerful tool for detecting and correcting label errors in language datasets



Human-originated noise

- Many datasets are crowdsourced -〉 human annotation errors or consistency..
- Selecting one of those label would be more realistic and challenging error



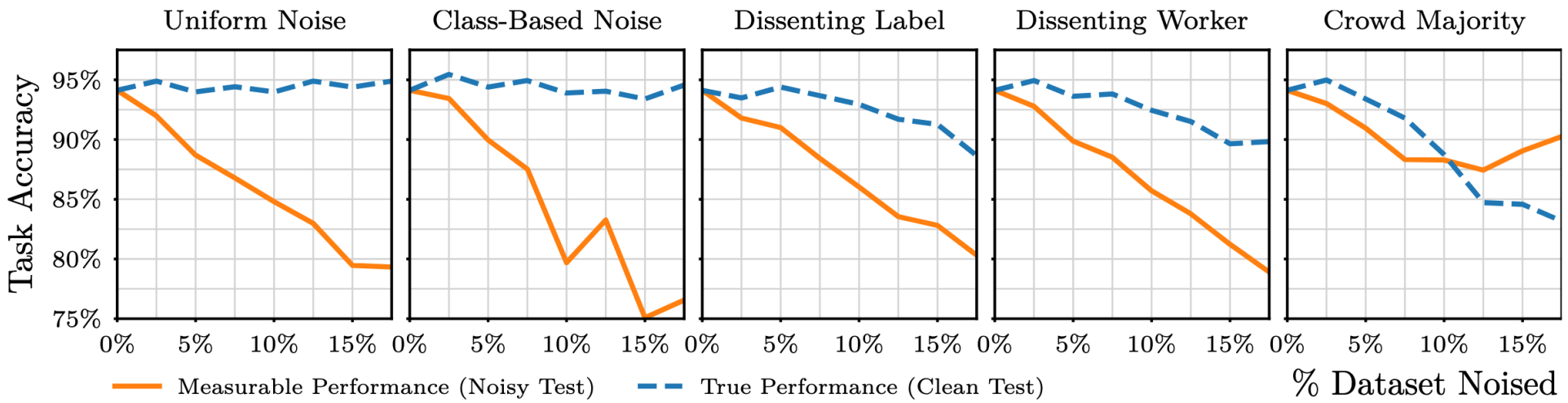
Human-originated noise

- Many datasets are crowdsourced -〉 human annotation errors or consistency…
- Selecting one of those label would be more realistic and challenging error

Dataset	Text	Label	Sentiment
IMDB	It is really unfortunate that a movie so well produced turns out to be such a disappointment. I thought this was full of (silly) cliches. It had all sorts of differences that it tried to tie together (not a bad thing in itself) but the result is at best awkward, but in fact ridiculous–too many clashes that wouldn't really happen. Then the end of the movie–the last 10 minutes–ruined all the rest. At first I thought Xavier was OK but with retrospect I think he was pretty bad. And that's all really too bad, because technically it was really good, and the soundtrack was great too. So the form was good, but the content pretty horrible.	~~Positive~~	Negative
IMDB	The ending made my heart jump up into my throat. I proceeded to leave the movie theater a little jittery. After all, it was nearly midnight. The movie was better than I expected. I don't know why it didn't last very long in the theaters or make as much money as anticipated. Definitely would recommend.	~~Negative~~	Positive
Amazon	The new design only has a thin layer of cellulose sponge material. It will not last as long. Already showing signs of wearing out. The picture does not represent the item received.	~~Neutral~~	Negative

Robustness does not generalize to human-originated noise

- Models may be robust to uniform and class-dependent noise
- PLM models not robust to human-originated noise



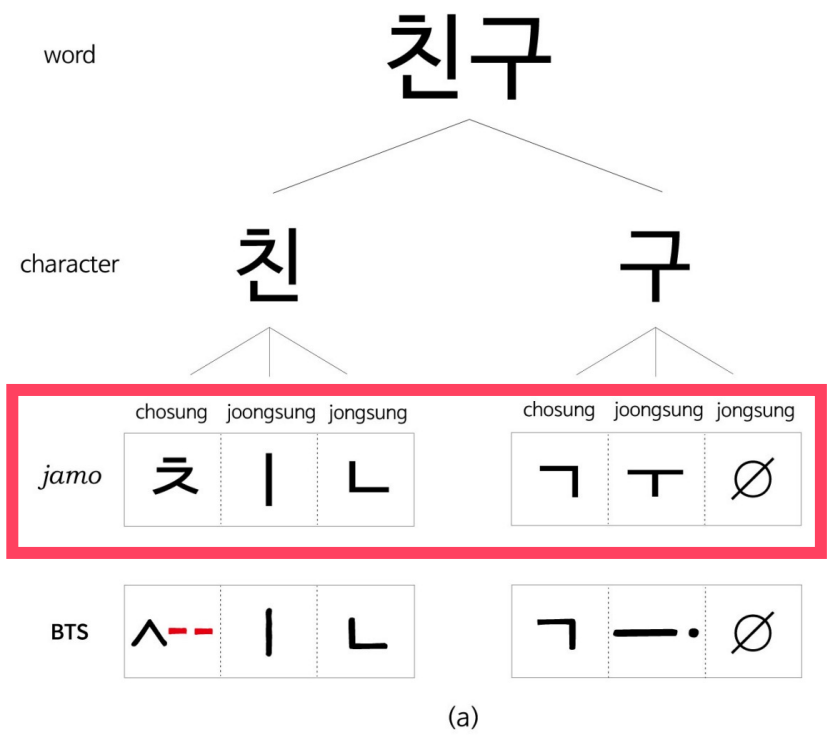
- Dissenting label: method replaces final labels with disagreeing labels at random, simulating imperfect quality control.
- Dissenting worker : select one annotator at random, apply all of their labels which disagree with final labels, and repeat until reaching the target noise rate.

Break it Down into BTS:
Basic, Tiniest Subword Units for Korean

Nayeon Kim[1*], **Jun-Hyung Park**[1*], **Joon-Young Choi**[2],
Eojin Jeon[2], **Youjin Kang**[1], **SangKeun Lee**[1,2]
[1]Department of Computer Science and Engineering [2]Department of Artificial Intelligence
Korea University, Seoul, Republic of Korea
{lilian1208, irish07, johnjames, skdlcm456, yjkang10, yalphy}@korea.ac.kr

Problem Statement

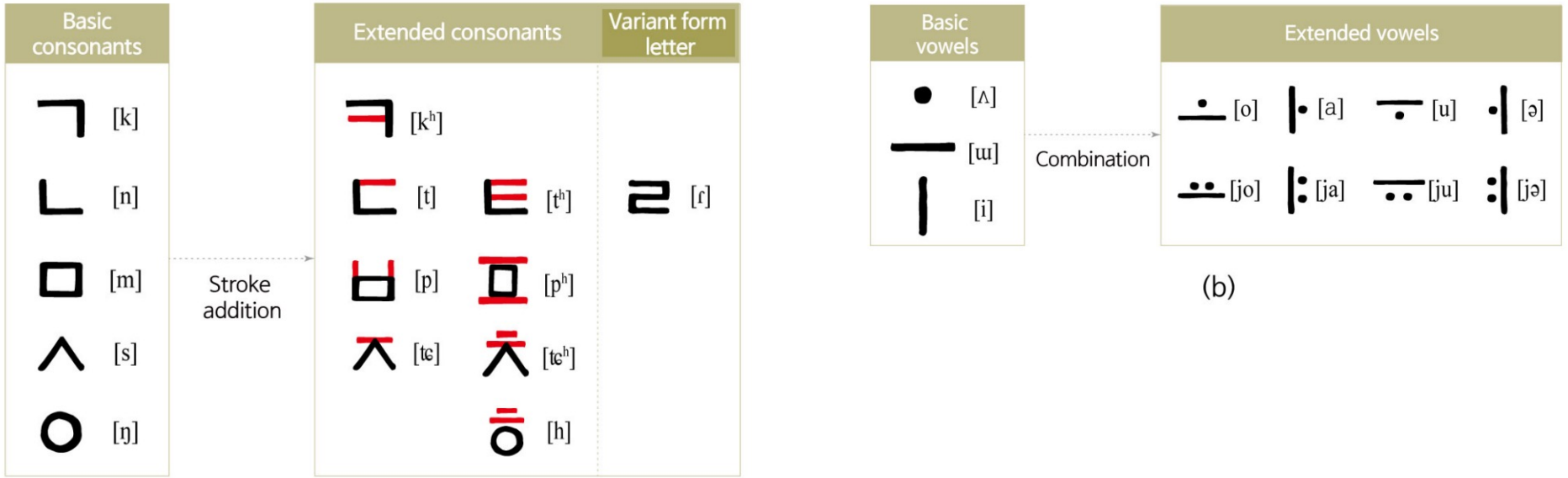
- 어떻게 언어적 지식 (linguistic knowledge) 을 사용하여 한국어 워드 임베딩의 품질을 개선 할 수 있을까?
- Hierarchical structure of the Korean word “친구”



(a)

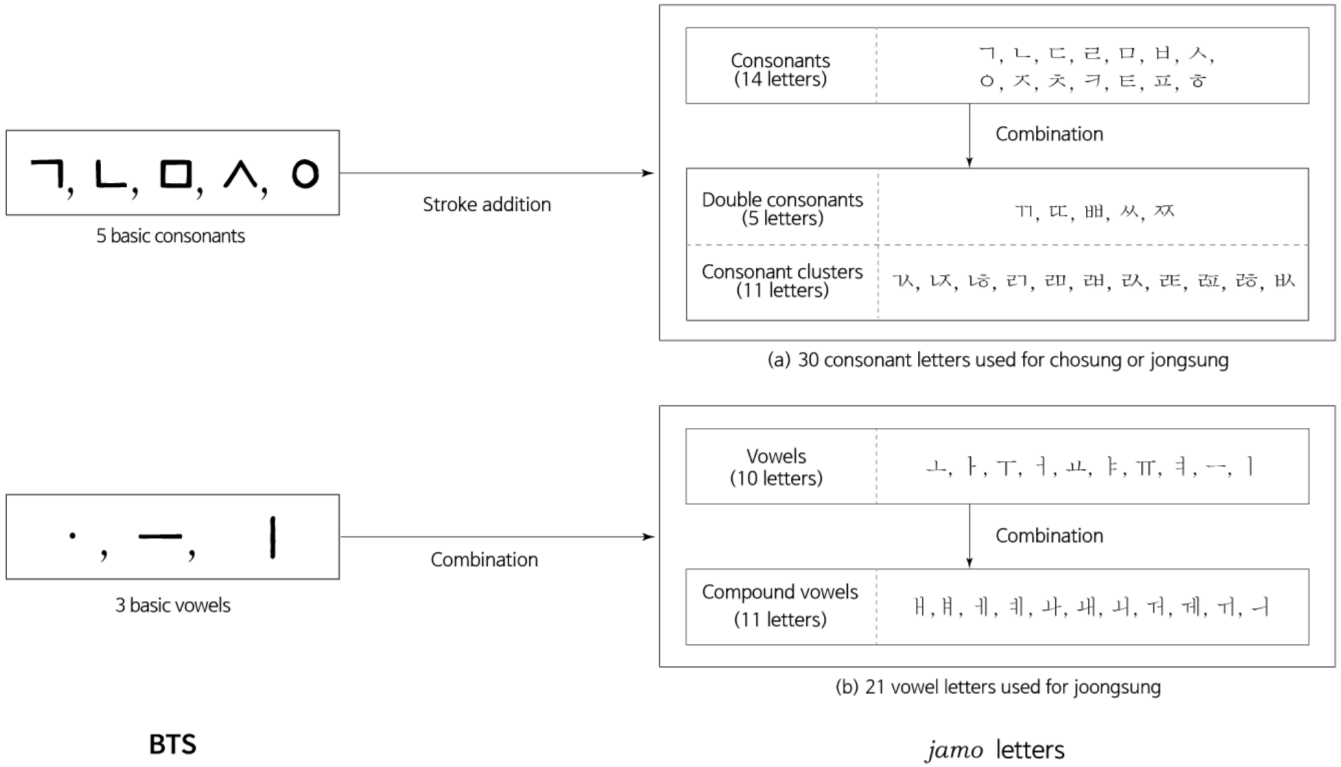
Motivation

- Hanguel involves five basic consonant(자음) and three basic vowels (모음)



Motivation

- Basic, Tiniest Subword Units for Korean
- BTS units comprise eight basic units
- Each of BTS units is n-stroke letter defined as an atomic unit



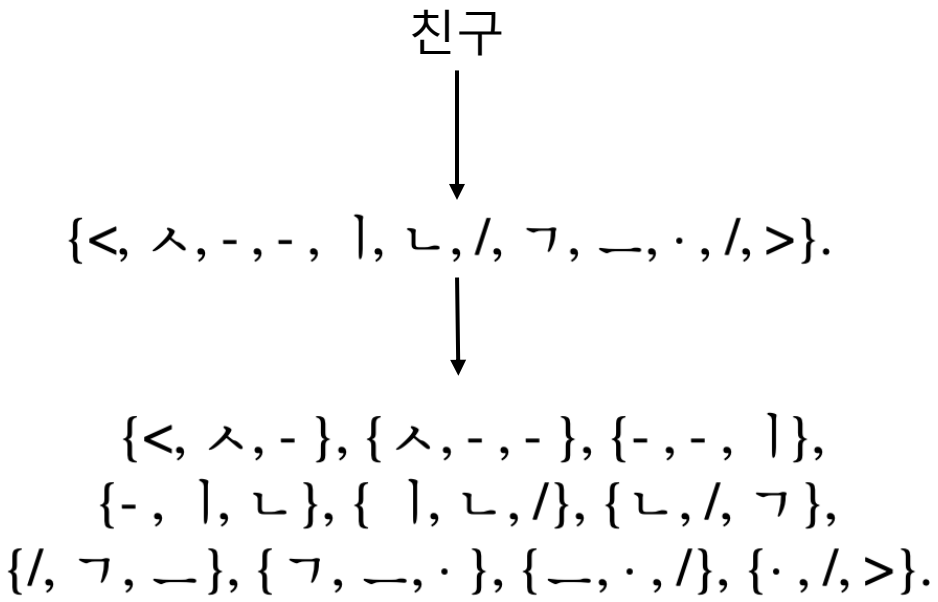
Method

- Example of subword decomposition

Decomposition Level	Subword Sequence
word	친구
character	친, 구
jamo (Park et al., 2018)	ㅊ, ㅣ, ㄴ, ㄱ, ㅜ
stroke (ours)	ㅅ, -, -, ㅣ, ㄴ, ㄱ, ㅜ
cji (ours)	ㅊ, ㅣ, ㄴ, ㄱ, ㅡ, ·
BTS (ours)	ㅅ, -, -, ㅣ, ㄴ, ㄱ, ㅡ, ·

Table 1: Examples of subword decomposition for the word '친구$_{\text{friend}}$' sorted by level of decomposition units

- Extracting BTS n-grams of "친구"
- Assigning vectors to each n-gram



Experiments

- **SISG (BTS)**: BTS의 subword 분절 단위를 사용하여 Skip-Gram embedding 수행

- Training Corpus
  - 2020 Newspaper Corpus
  - Korean Wikipedia
  - 21st Century Sejong Corpus

- Evaluation Tasks
  - Word Analogy
  - Word Similarity
  - Sentimental Analysis (Naver movie review)

Result: Word Analogy

- evaluate the semantic and syntactic features of word vectors on the Korean word
- A is to B as C is to D, What is D?

Model	Semantic					Syntactic					Sem. Avg.	Syn. Avg.	Avg.
	Capt	Gend	Name	Lang	Misc	Case	Tense	Voice	Form	Honr			
SG	0.463	0.531	0.585	0.435	0.644	0.533	0.612	0.543	0.677	0.538	0.532	0.581	0.556
SISG(ch)	0.417	0.460	0.554	0.374	0.561	0.234	0.472	0.456	0.545	0.357	0.473	0.413	0.443
SISG(jm)	0.413	0.430	0.510	0.346	0.557	0.164	0.351	0.364	0.415	0.297	0.451	0.318	0.385
SISG(ch4+jm)	0.402	0.432	0.506	0.337	0.556	0.152	0.346	0.361	0.404	0.294	0.447	0.311	0.379
SISG(ch6+jm)	0.404	0.430	0.502	0.337	0.556	0.151	0.345	0.364	0.400	0.295	0.446	0.311	0.378
SISG(stroke)	0.347	0.368	0.448	0.309	0.481	0.154	0.324	0.352	0.380	**0.260**	0.391	0.294	0.342
SISG(cji)	0.347	0.368	0.447	0.312	0.485	0.156	0.321	0.355	0.374	0.268	0.392	0.295	0.343
SISG(**BTS**)	**0.342**	**0.360**	**0.440**	**0.306**	**0.473**	**0.151**	**0.319**	**0.348**	**0.370**	0.267	**0.384**	**0.291**	**0.338**
SISG(jm+stroke)	0.343	0.360	0.446	**0.303**	0.476	0.151	0.316	0.351	**0.363**	0.256	0.386	0.287	**0.337**
SISG(jm+cji)	0.350	0.383	0.444	0.312	0.497	0.152	0.319	0.349	0.387	0.265	0.397	0.294	0.346
SISG(jm+**BTS**)	**0.339**	0.362	**0.443**	0.305	**0.475**	0.153	0.327	0.355	0.369	0.264	**0.385**	0.294	0.339
SISG(ch4+stroke)	0.346	**0.358**	0.451	**0.303**	0.477	0.153	0.319	0.352	0.372	0.260	0.387	0.291	0.339
SISG(ch4+cji)	0.352	0.382	0.445	0.316	0.502	0.153	0.322	0.347	0.389	0.266	0.399	0.296	0.347
SISG(ch4+**BTS**)	0.346	0.389	0.444	0.311	0.499	0.159	0.338	0.358	0.407	0.269	0.398	0.306	0.352
SISG(ch6+stroke)	0.348	0.381	0.447	0.309	0.492	0.153	0.328	0.352	0.391	0.266	0.395	0.298	0.347
SISG(ch6+cji)	0.349	0.376	0.452	0.312	0.486	0.158	0.328	0.362	0.384	0.273	0.395	0.301	0.348
SISG(ch6+**BTS**)	0.348	0.372	0.447	0.307	0.490	**0.146**	**0.314**	**0.337**	0.378	**0.253**	0.393	**0.286**	0.339

Result: Word Similarity

- evaluate the trained word vectors on how well they formulate the relationship between words

Model	Similarity
SG	0.591
SISG(ch)	0.665
SISG(jm)	0.675
SISG(ch4+jm)	0.687
SISG(ch6+jm)	0.684
SISG(stroke)	0.703
SISG(cji)	**0.707**
SISG(**BTS**)	**0.707**

Result: Sentimental Analysis

- evaluate the trained word vectors on how well they formulate the relationship between words

Model	Acc.	Prc.	Rec.	F1
SG	78.07	0.818	0.738	0.776
SISG(ch)	81.03	0.876	0.732	0.797
SISG(jm)	81.83	0.865	0.762	0.810
SISG(stroke)	82.44	0.878	0.758	0.814
SISG(cji)	**82.50**	0.862	0.781	**0.820**
SISG(**BTS**)	82.18	0.843	**0.798**	**0.820**

Result: Nearest Neighbor wood

- top3 nearest neighbors for Korean words
- better capability in identifying the meanings of words containing typos

Query	SG		SISG(ch)		SISG(jm)		SISG(**BTS**)	
	NN words	Equiv	NN words	Equiv	NN words	Equiv	NN words	Equiv
케잌	버터크림	✗	프랄린	✗	케익	✓	케익	✓
	버터크림을	✗	태슬	✗	케일	✗	케익을	✓
	카나페	✗	피넛	✗	케인	✗	케이크$_{\text{cake}}$	✓
찐한	사랑법이	✗	틀막	✗	찐득한	✗	찐득한	✗
	애절함	✗	3MC의	✗	찐찌	✗	진한$_{\text{strong}}$	✓
	애틋하면서도	✗	잔망	✗	찐	✗	찐	✗
웬지	괜스레	✗	어쩐지	✗	어쩐지	✗	웬일	✗
	허전하고	✗	웬일인지	✗	왠지$_{\text{for some reason}}$	✓	웬일이지	✗
	초조하고	✗	웬일이지	✗	웬일인지	✗	웬일이니	✗
뒤치닥거리	*n/a*	*n/a*	푸닥거리	✗	푸닥거리	✗	뒤치다꺼리$_{\text{cover for}}$	✓
	n/a	*n/a*	복닥거리는	✗	푸닥거리를	✗	뒤치다꺼리를	✓
	n/a	*n/a*	노닥거리는	✗	노닥거리고	✗	푸닥거리	✗

Conclusion

- 한글 창제 원리에 기반한 임베딩 방법론 제시
- Problem statement와 motivation, method는 정말 중요한 요소
- 너무 어려운 Problem statement?

LM-Debugger:

An Interactive Tool for Inspection and Intervention in Transformer-Based Language Models

Mor Geva[1] **Avi Caciularu**[2,*] **Guy Dar**[3] **Paul Roit**[2] **Shoval Sadde**[1]
Micah Shlain[1] **Bar Tamir**[4] **Yoav Goldberg**[1,2]

[1]Allen Institute for AI [2]Bar-Ilan University
[3]Tel Aviv University [4]The Hebrew University of Jerusalem

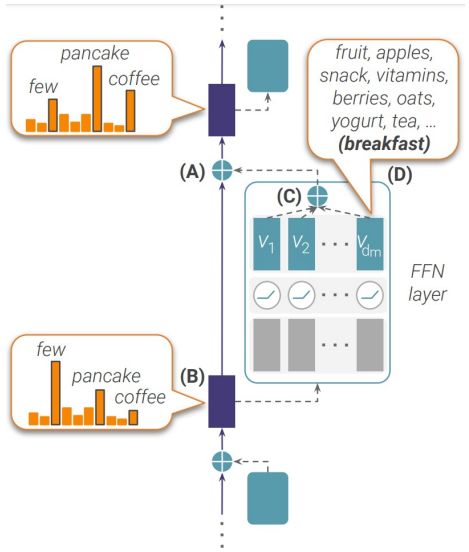
morp@allenai.org

Problem Statement

- Transformer-based language models (LMs) are widely used in NLP
- but their internal prediction construction process is **opaque** difficult to understand

  → challenging for end-users to understand

  → model makes specific predictions and for developers to debug or fix model behavior

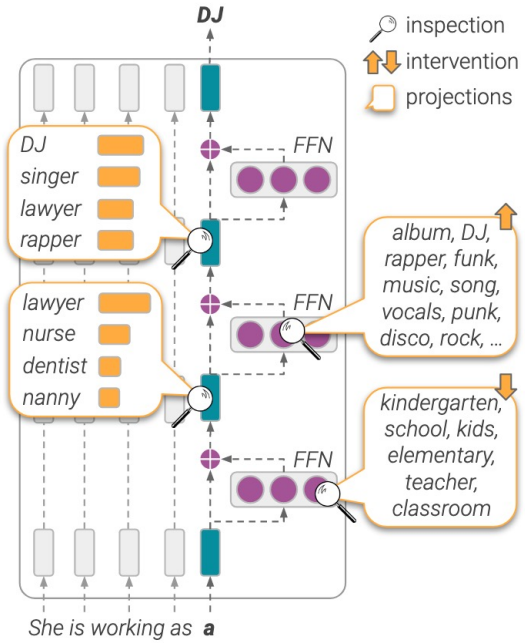


<Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space>

Method

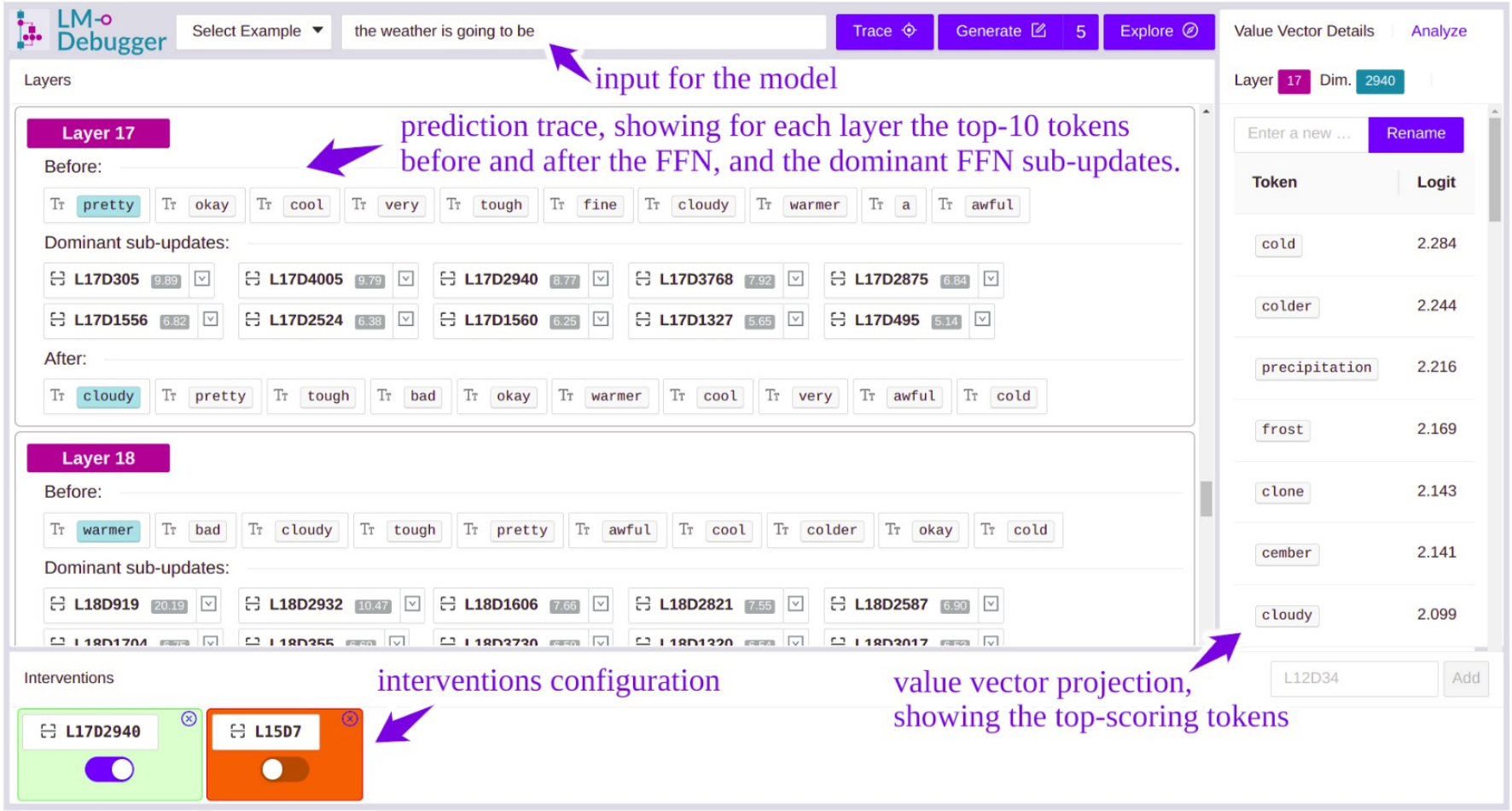
LM-Debugger

- LM-Debugger: a tool for inspecting and intervening in transformer LM predictions.
- LM-Debugger projects the token representation before and after the feedforward network (FFN) updates
- intervening in the prediction by changing the weights of specific subupdates



Method

LM-Debugger



Conclusion

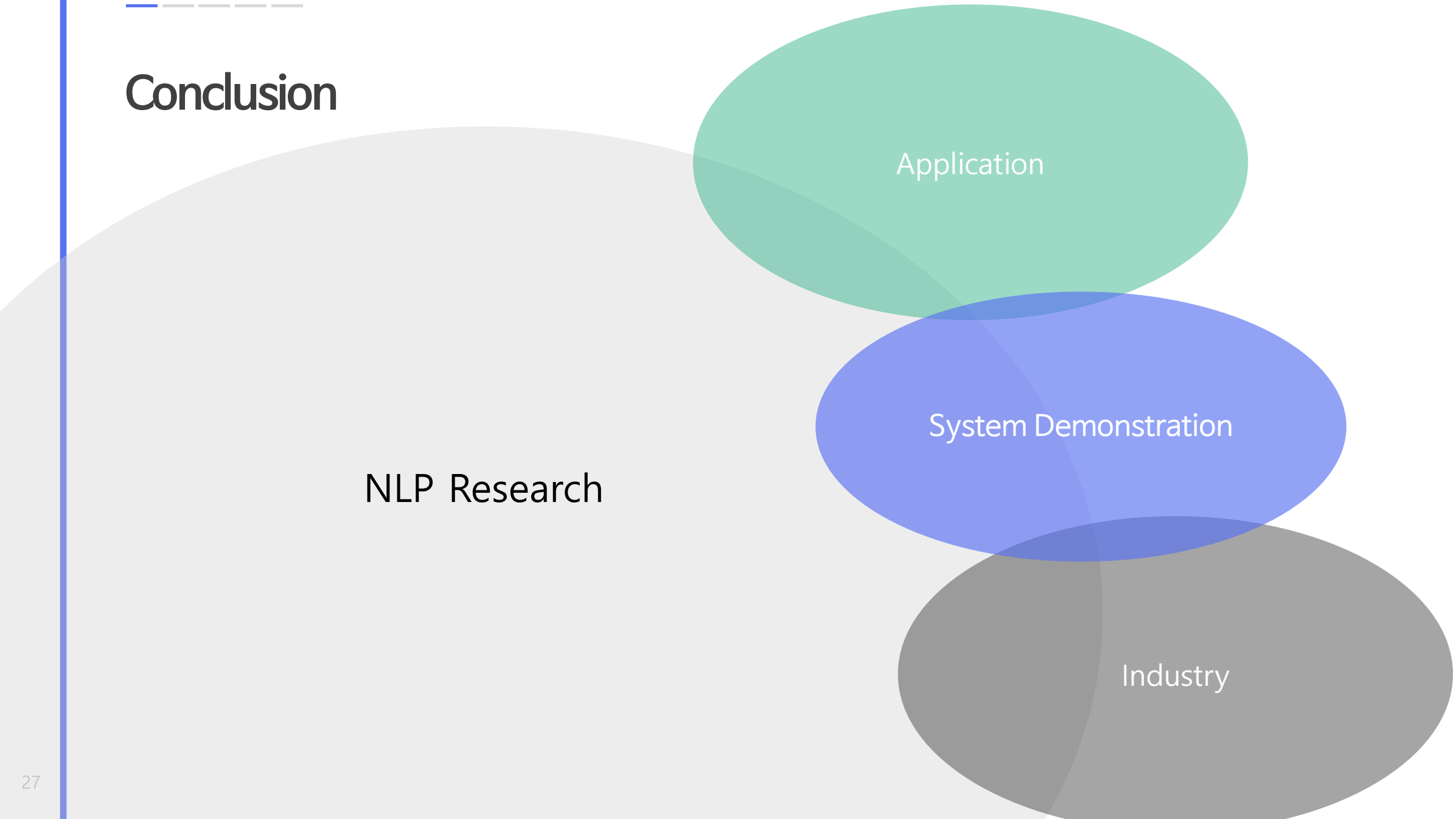
- Detecting Label Errors by using Pre-Trained Language Models
  - Real world에서 문제가 되는 label error
- Break it Down into BTS: Basic, Tiniest Subword Units for Korean
  - 한글 창제 원리에 기반한 임베딩 방법론 제시
  - 너무 어려운 Problem statement?
- LM-Debugger
  - 실험 X 이지만, 매력적인 NLP Researcher 들에게 지적되는 문제를 statement
- **결론**
  - Problem statement와 motivation, method는 정말 중요한 요소
  - Research와 application을 아우르는 연구는 정말 어려운 요소

Conclusion

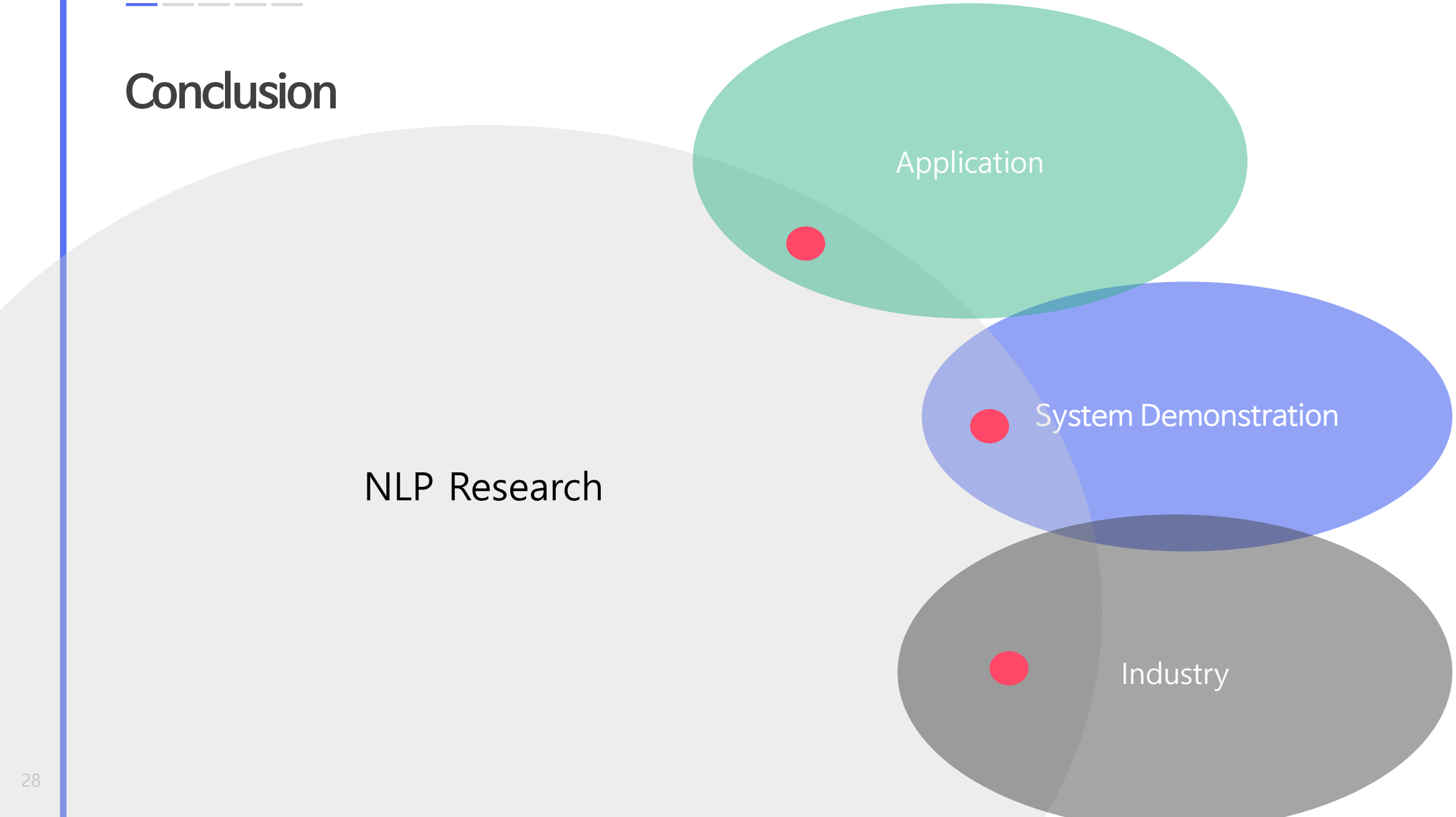
“내가 재미 없다고 해서 그 영화가 미학적 가치가 없는 것은 아니다”

– 영화 평론가 이동진, 기생충 인터뷰 중 –

Conclusion



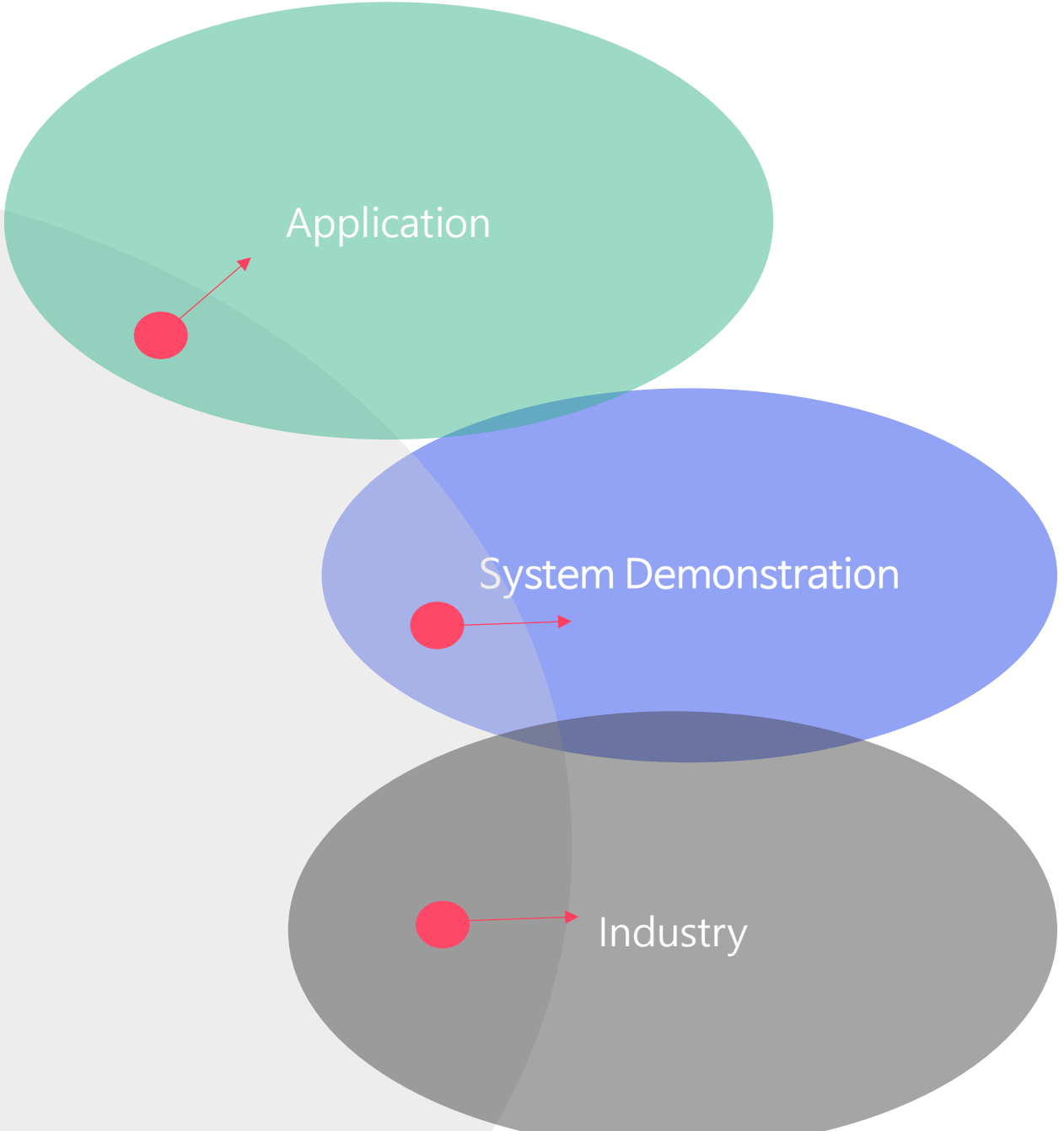
Conclusion



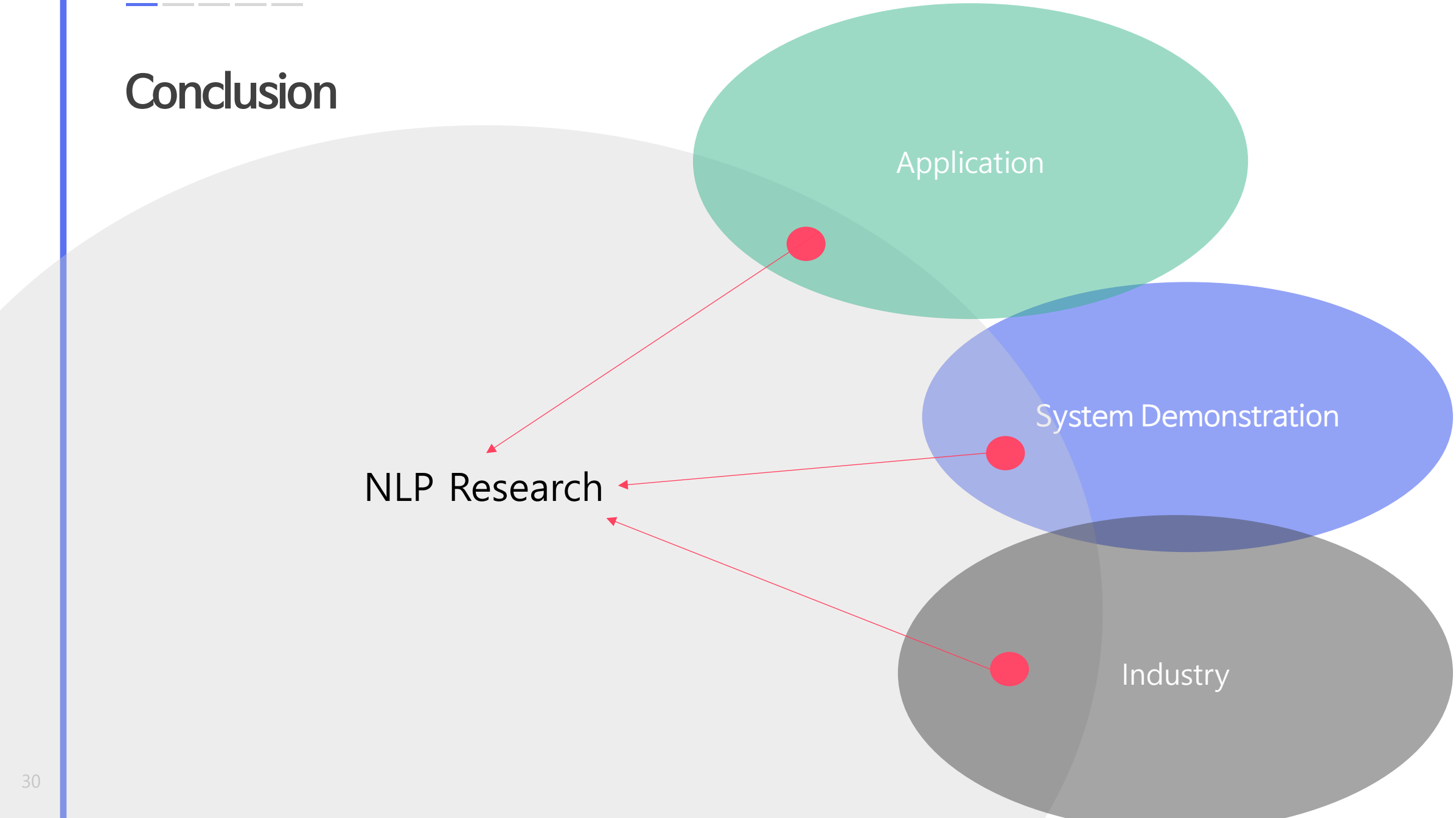
NLP Research

Conclusion

NLP Research



Conclusion



Human-originated noise is realistic and challenging

- Uniform and class-based noise produce high and distinctive losses
- human-originated noise is widely distributed