Indirect Supervision from Natural Language Inference

Indirectly Supervised Natural Language Processing (Part I)

Wenpeng Yin
Department of Computer Science and Engineering
Penn State University

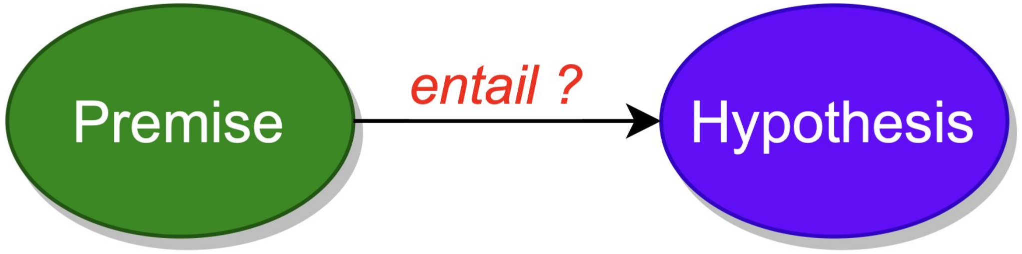
July 2023

ACL Tutorials

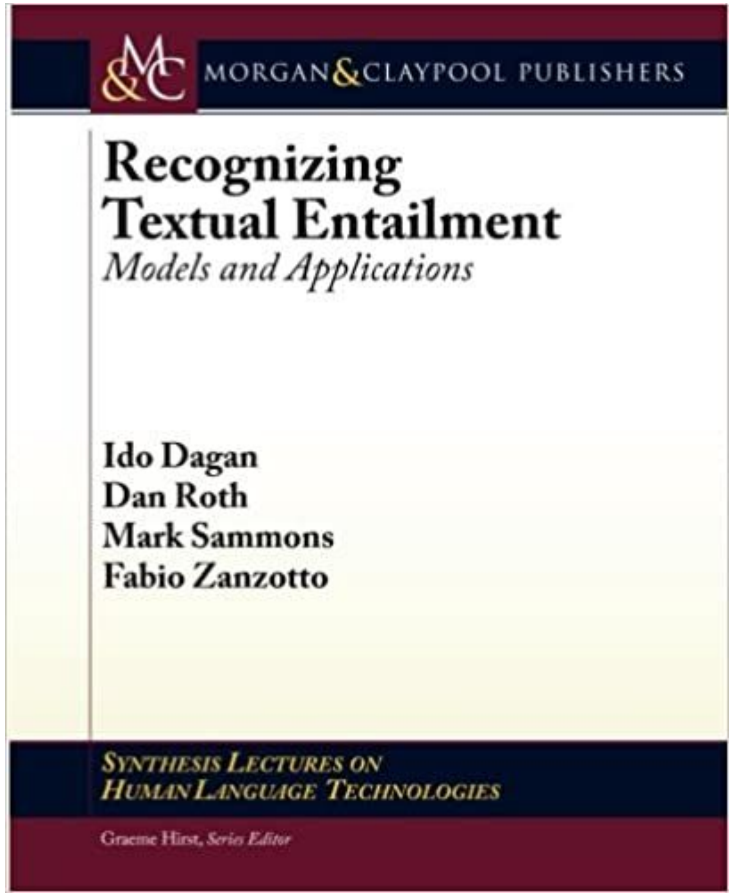
Indirectly Supervised Natural Language Processing

Natural Language Inference for NLP

- Textual Entailment (Dagan et al., 2006)

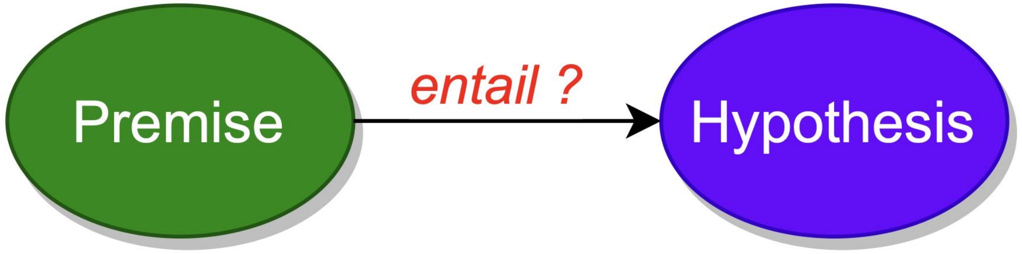


- Textual entailment, a unified inference framework for NLP
  - zero-shot text classification (Yin et al., 2019)
  - summarization (Falke et al., 2019)
  - QA & Coreference (Yin et al., 2020)
  - relation extraction (Xia et al., 2021)
  - entity typing (Li et al., 2022)
  - …

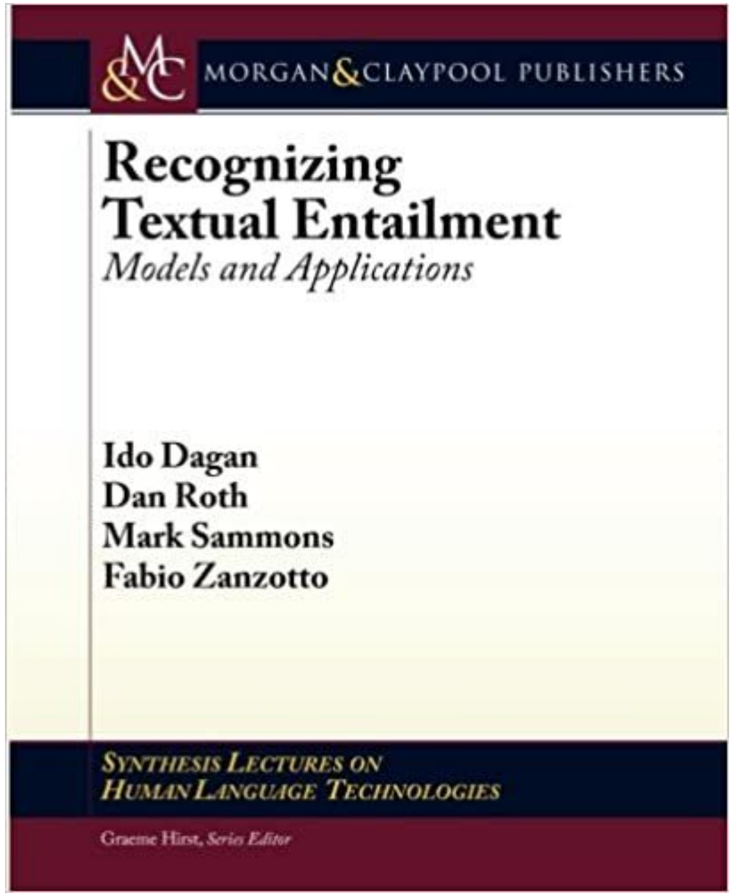


Natural Language Inference for NLP

- ❑ Textual Entailment (Dagan et al., 2006)



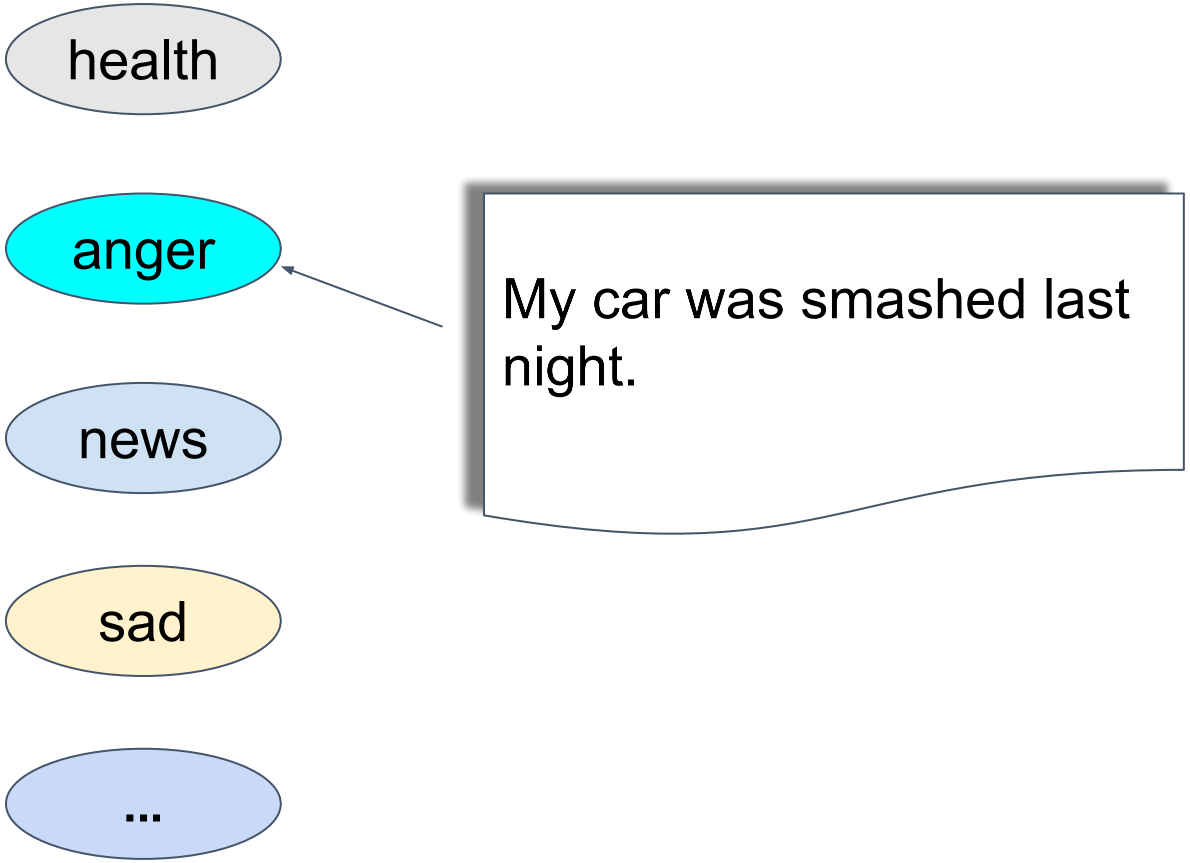
- ❑ Textual entailment, a unified inference framework for NLP
  - ❑ zero-shot text classification (Yin et al., 2019)
  - ❑ summarization (Falke et al., 2019)
  - ❑ QA & Coreference (Yin et al., 2020)
  - ❑ relation extraction (Xia et al., 2021)
  - ❑ entity typing (Li et al., 2022)
  - ❑ …



- ❑ "Textual Entailment" was referred to as "<u>Natural Language Inference (NLI)</u>" (Bowman et al., 2015)

NLI, unified inference in various NLP tasks

Example task: Text Classification



NLI, unified inference in various NLP tasks

Example task: Text Classification

health

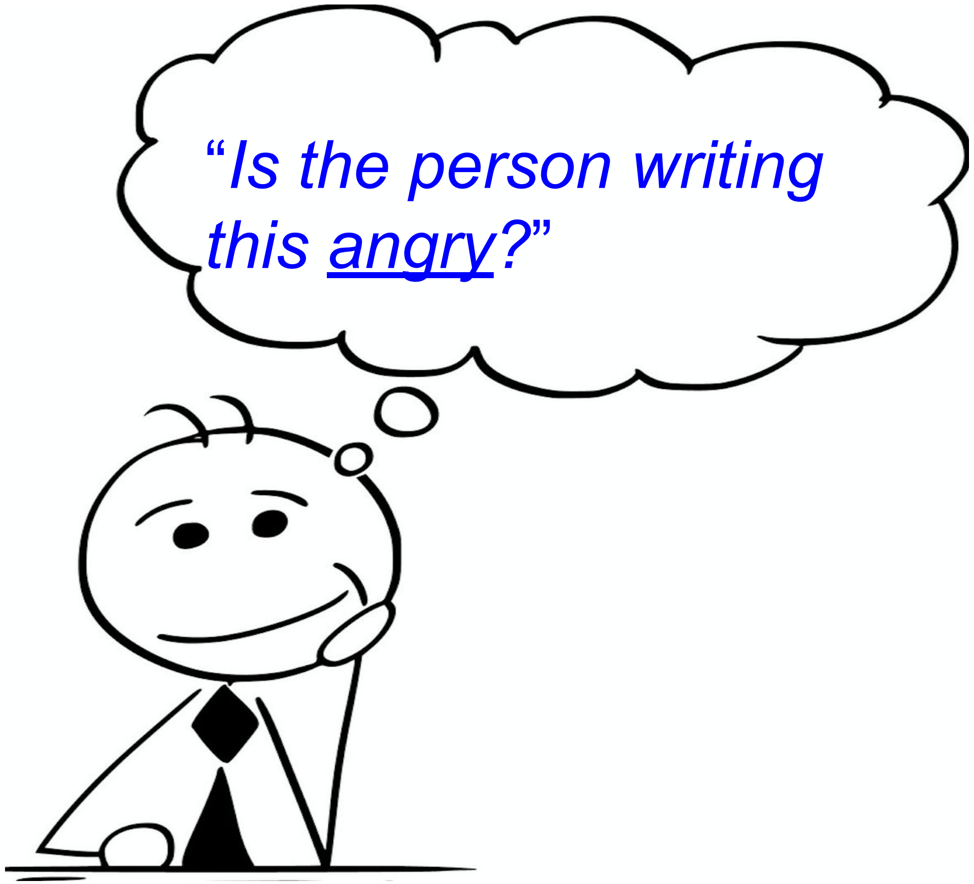
anger

news

sad

...

My car was smashed last night.



Humans implicitly build a hypothesis then infer its truth value.

Example task: Text Classification

My car was smashed last night.

The person is angry.

The person is not angry.

The classification problem is **inference (text → hypothesis)**

Example task: Question Answering

The road to Grandpa's house was long and winding. [...]. Finally, Jimmy arrived at Grandpa's house and knocked. Grandpa answered the door with a smile and welcomed Jimmy inside. They sat leisurely by the fire and talked about the insects. They watched the lightning bugs light up as night came.

Where do Jimmy and his Grandpa sit?

A) On insects
B) Outside
C) By the fire
D) On the path

NLI, unified inference in various NLP tasks

Example task: Question Answering

The road to Grandpa's house was long and winding. [...]. Finally, Jimmy arrived at Grandpa's house and knocked. Grandpa answered the door with a smile and welcomed Jimmy inside. They sat leisurely by the fire and talked about the insects. They watched the lightning bugs light up as night came.

Where do Jimmy and his Grandpa sit?

A) On insects
B) Outside
C) By the fire
D) On the path

Because "by the fire" and "sat" co-occur in the same sentence?

NLI, unified inference in various NLP tasks

Example task: Question Answering

The road to Grandpa's house was long and winding. [...]. Finally, Jimmy arrived at Grandpa's house and knocked. <u>Grandpa answered the door with a smile and welcomed Jimmy inside. They sat leisurely by the fire and talked about the insects</u>. They watched the lightning bugs light up as night came.

Where do Jimmy and his Grandpa sit?

A) On insects
B) Outside
C) <u>By the fire</u>
D) On the path

can infer the meaning

<u>"Jimmy and his Grandpa sit by the fire"</u>

NLI, unified inference in various NLP tasks

Example task: Question Answering (QA)

The road to Grandpa's house was long and winding. [...]. Finally, Jimmy arrived at Grandpa's house and knocked. Grandpa answered the door with a smile and welcomed Jimmy inside. They sat leisurely by the fire and talked about the insects. They watched the lightning bugs light up as night came.

Where do Jimmy and his Grandpa sit?

A) On insects
B) Outside
C) <u>By the fire</u>
D) On the path

QA is **inference (text → hypothesis)**

NLI, unified inference in various NLP tasks

Example task: Coreference Resolution

?

The **trophy** would not fit in the brown **suitcase** because <u>it</u> was too big.

?

NLI, unified inference in various NLP tasks

Example task: Coreference Resolution

?

The **trophy** would not fit in the brown **suitcase** because <u>it</u> was too big.

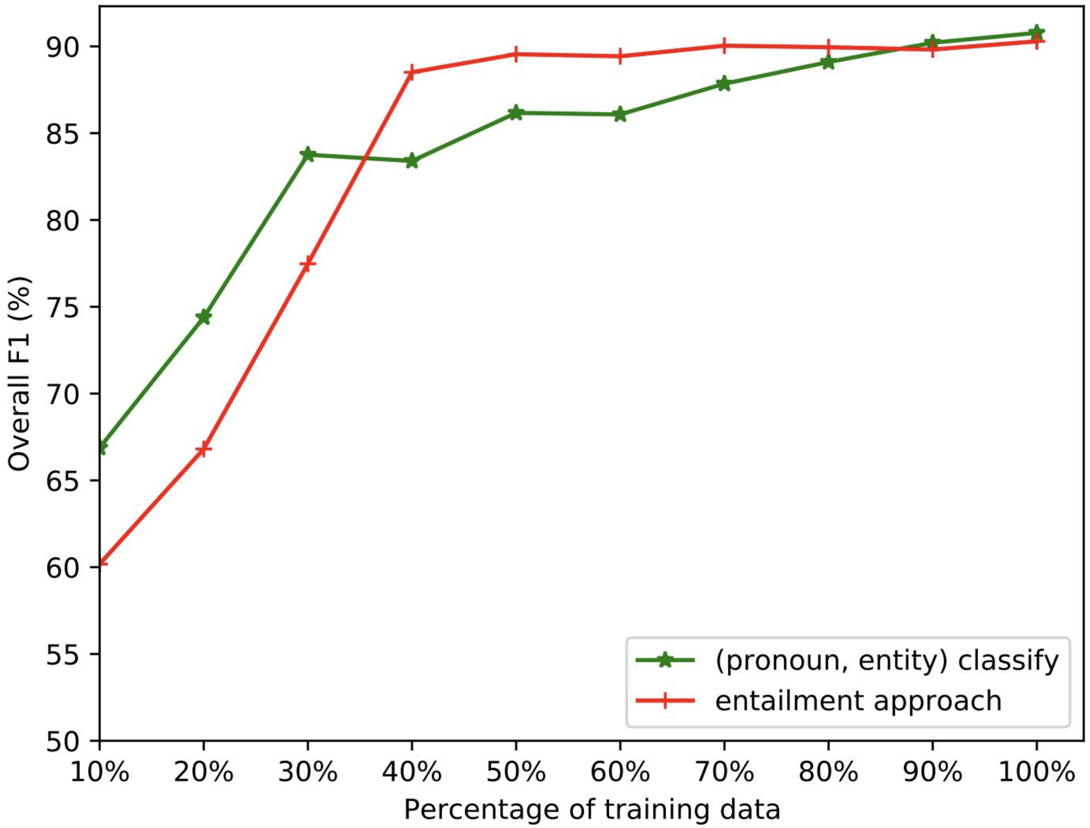
?

➔ The trophy would not fit in the brown suitcase because <u>**trophy**</u> was too big. (True)

➔ The trophy would not fit in the brown suitcase because <u>**suitcase**</u> was too big. (False)

Coreference resolution is **inference (text → hypothesis)**

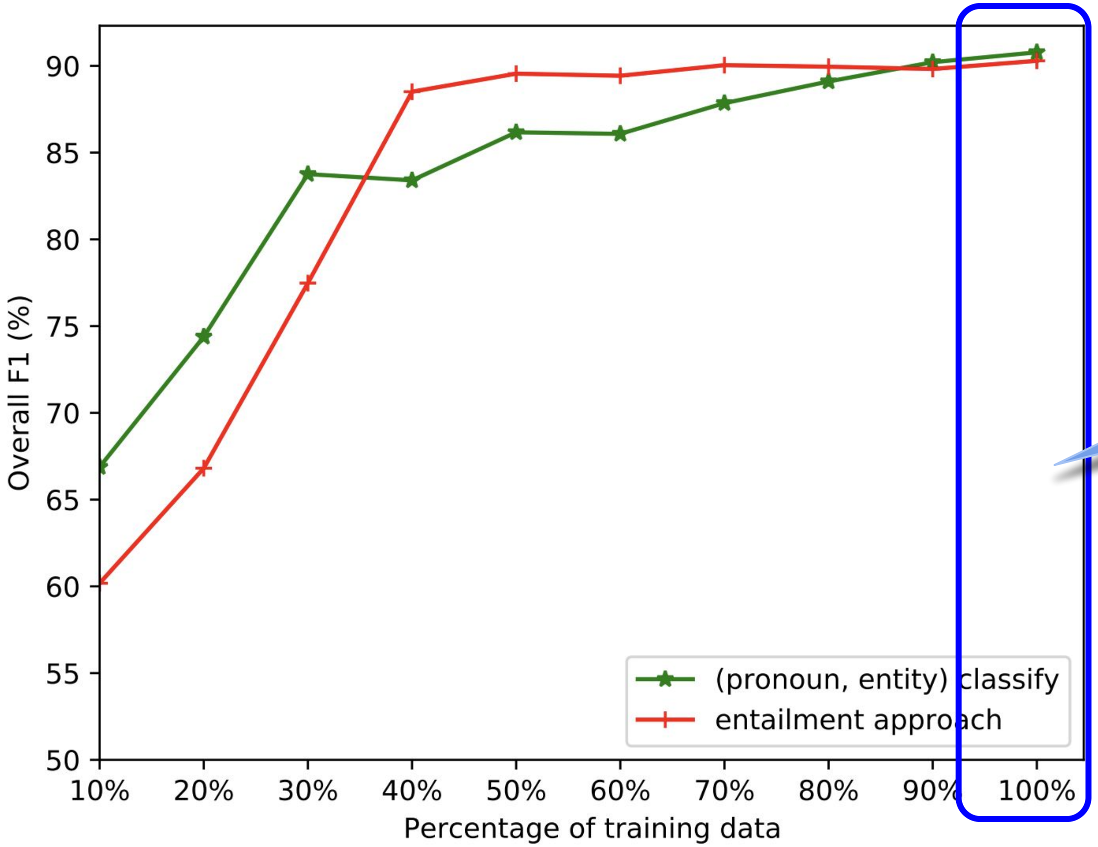
Why and when to convert NLP to NLI?



Task: conference resolution
Dataset: GAP (Webster et al., 2018)

Universal Natural Language Processing with Limited Annotations: Try Few-shot Textual Entailment as a Start (Yin et al., EMNLP'2020)

Why and when to convert NLP to NLI?

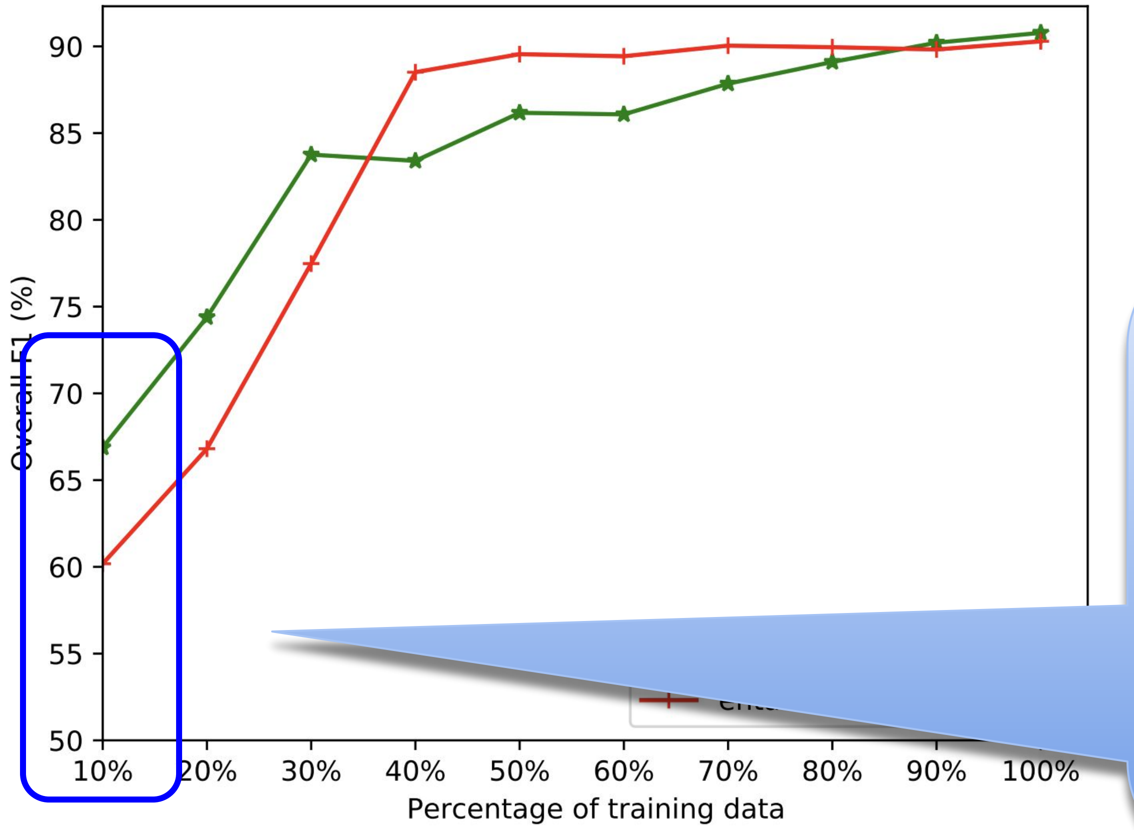


When an NLP task has rich annotations:
classical classifiers ≈ NLI

Task: conference resolution
Dataset: GAP (Webster et al., 2018)

Universal Natural Language Processing with Limited Annotations: Try Few-shot Textual Entailment as a Start (Yin et al., EMNLP'2020)

Why and when to convert NLP to NLI?



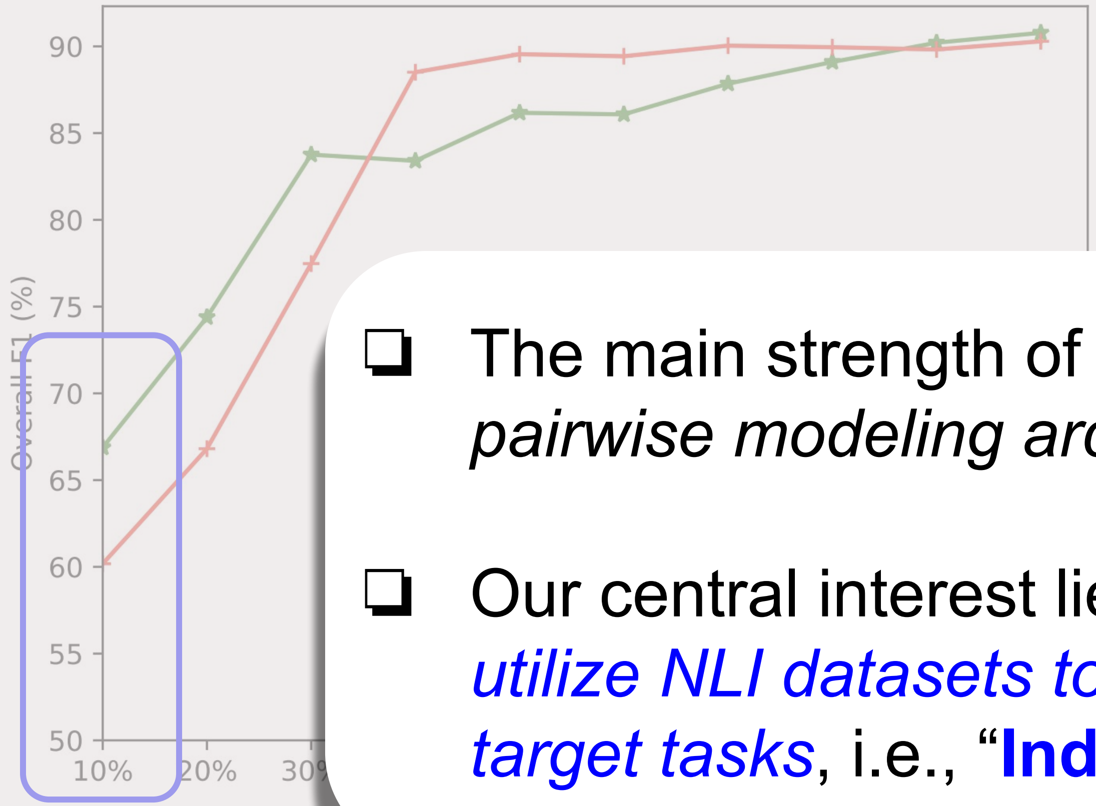
In reality, what we truly care about is "limited annotation"

- ❑ Both standard classifiers and NLI exhibit poor performance.
- ❑ NLI performs even worse because NLI is generally a more challenging task when the availability of labeled examples is severely limited.

Task: conference resolution
Dataset: GAP (Webster et al., 2018)

Universal Natural Language Processing with Limited Annotations: Try Few-shot Textual Entailment as a Start (Yin et al., EMNLP'2020)

Why and when to convert NLP to NLI?



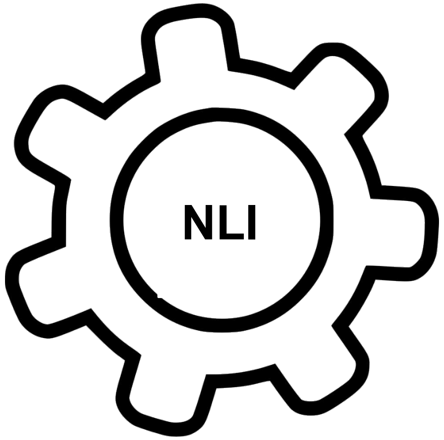
- ❑ The main strength of NLI is **NOT** its *pairwise modeling architecture*.
- ❑ Our central interest lies in its potential to *utilize NLI datasets to supervise different target tasks*, i.e., "**Indirect Supervision**"

Task: conference resolution
Dataset: GAP (Webster et al., 2018)

Universal Natural Language Processing with Limited Annotations: Try Few-shot Textual Entailment as a Start (Yin et al., EMNLP'2020)

NLI-based indirect supervision: outline

- ❑ Implementation & Applications

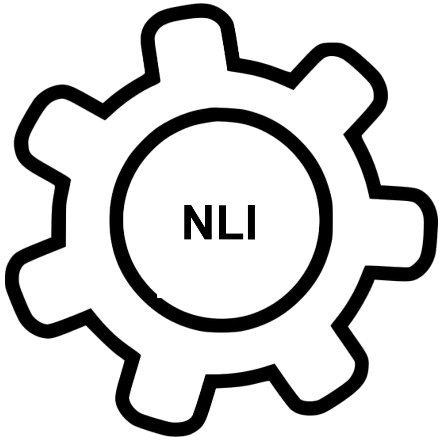


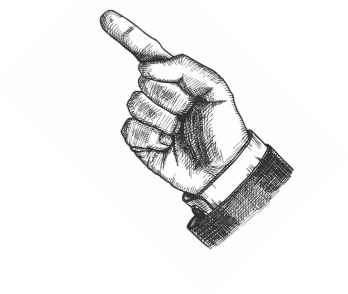
- ❑ Benefits

- ❑ Challenges & Solutions

NLI-based indirect supervision: outline

❑ Implementation & Applications

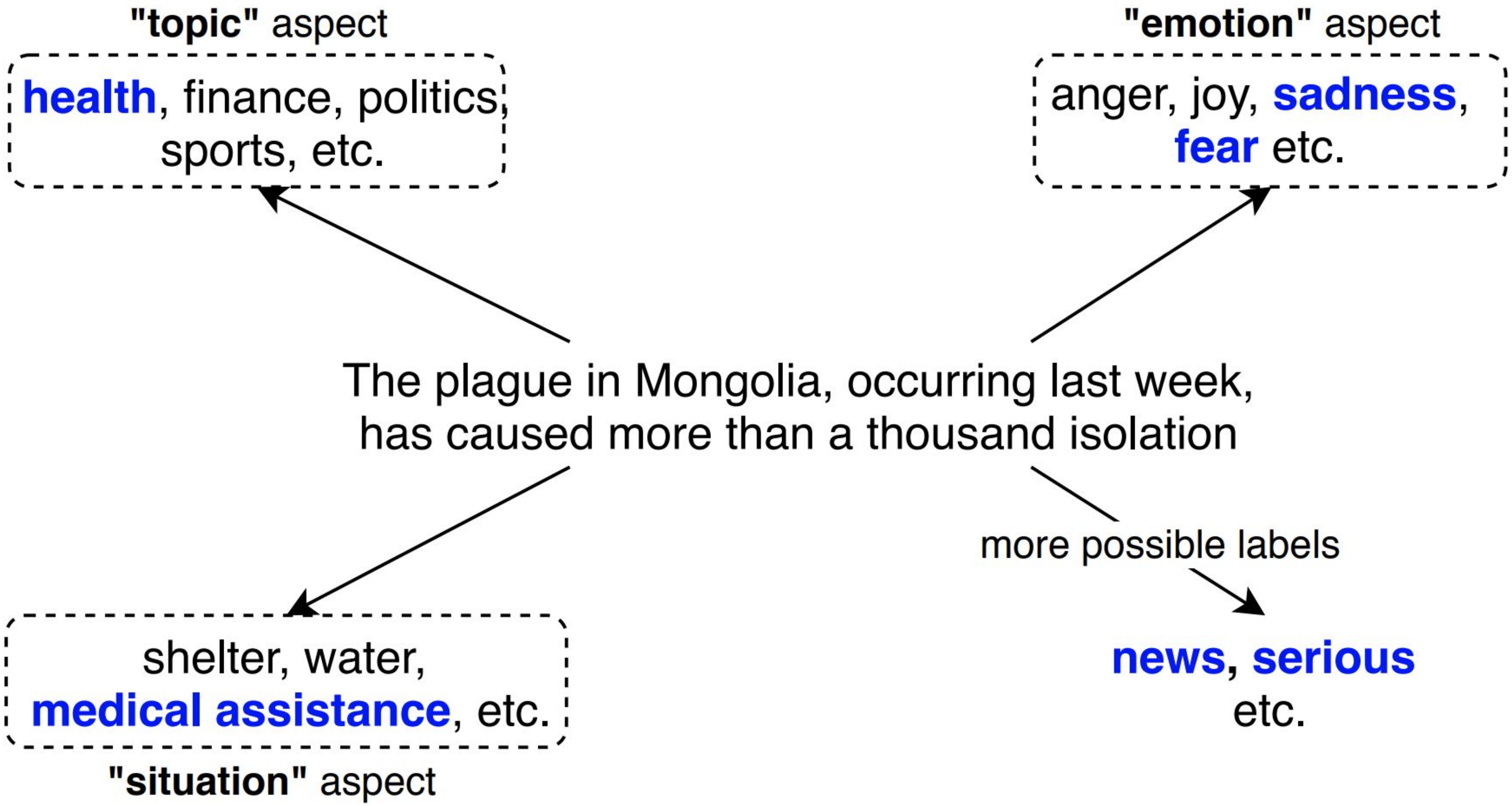




❑ Benefits

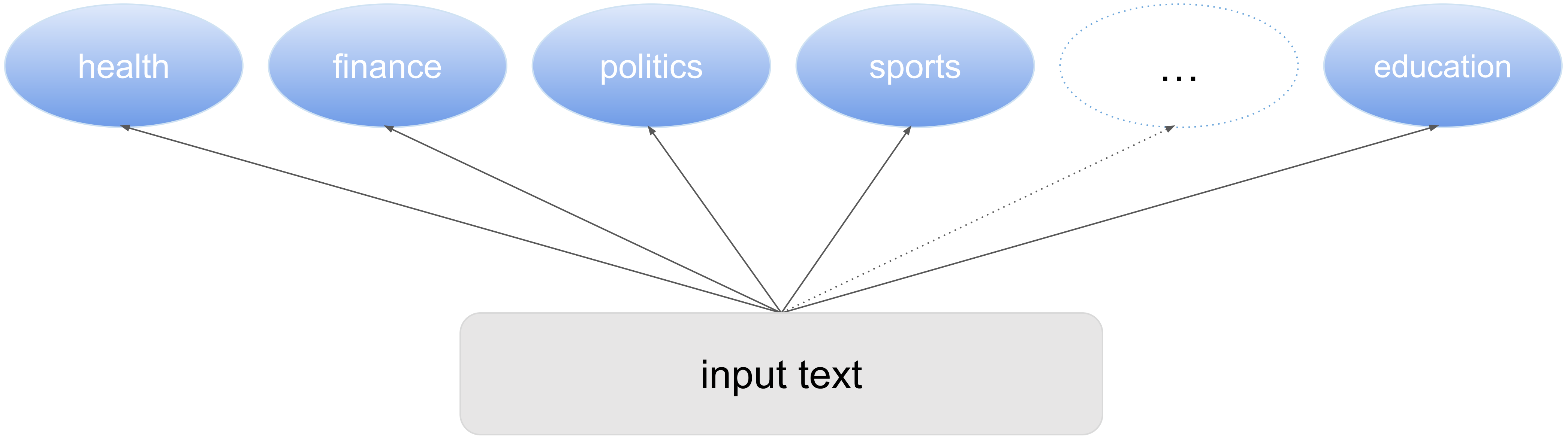
❑ Challenges & Solutions

Text classification task

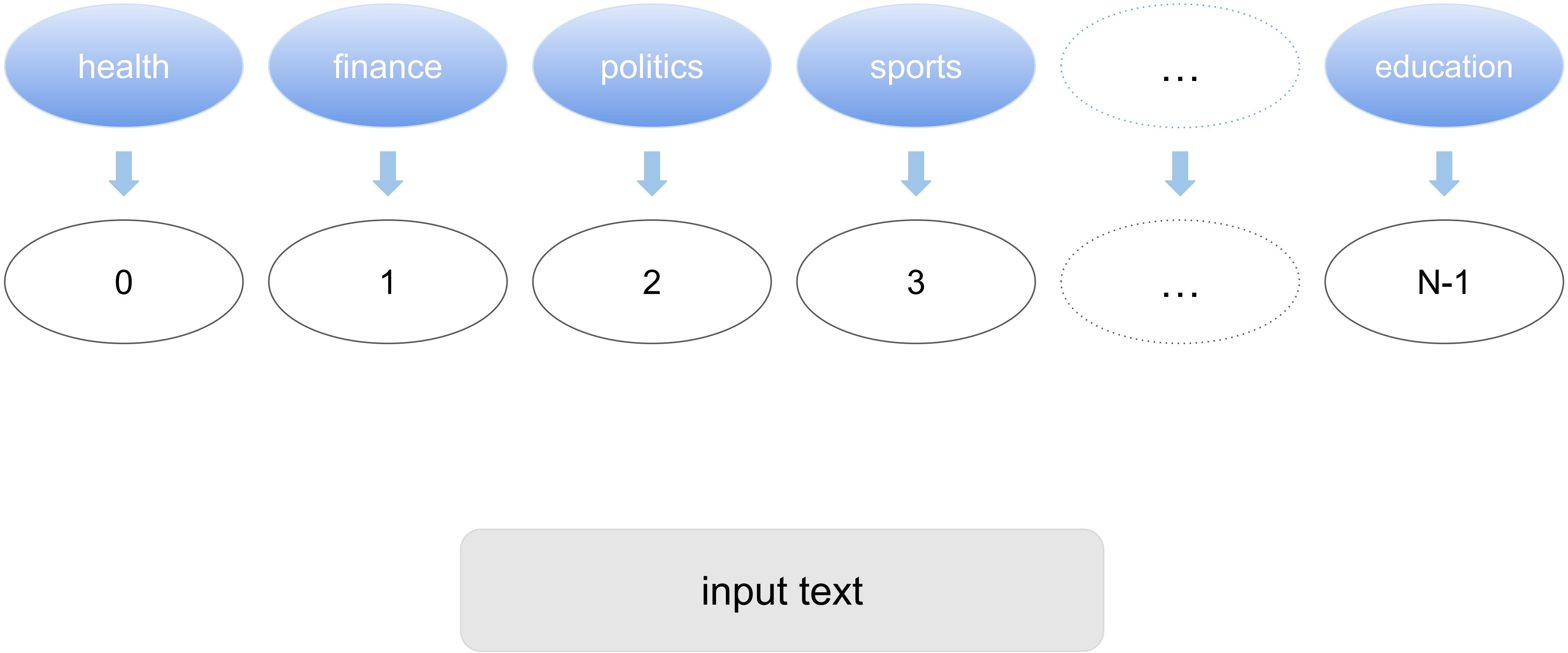


[Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach]() (Yin et al., 2019)

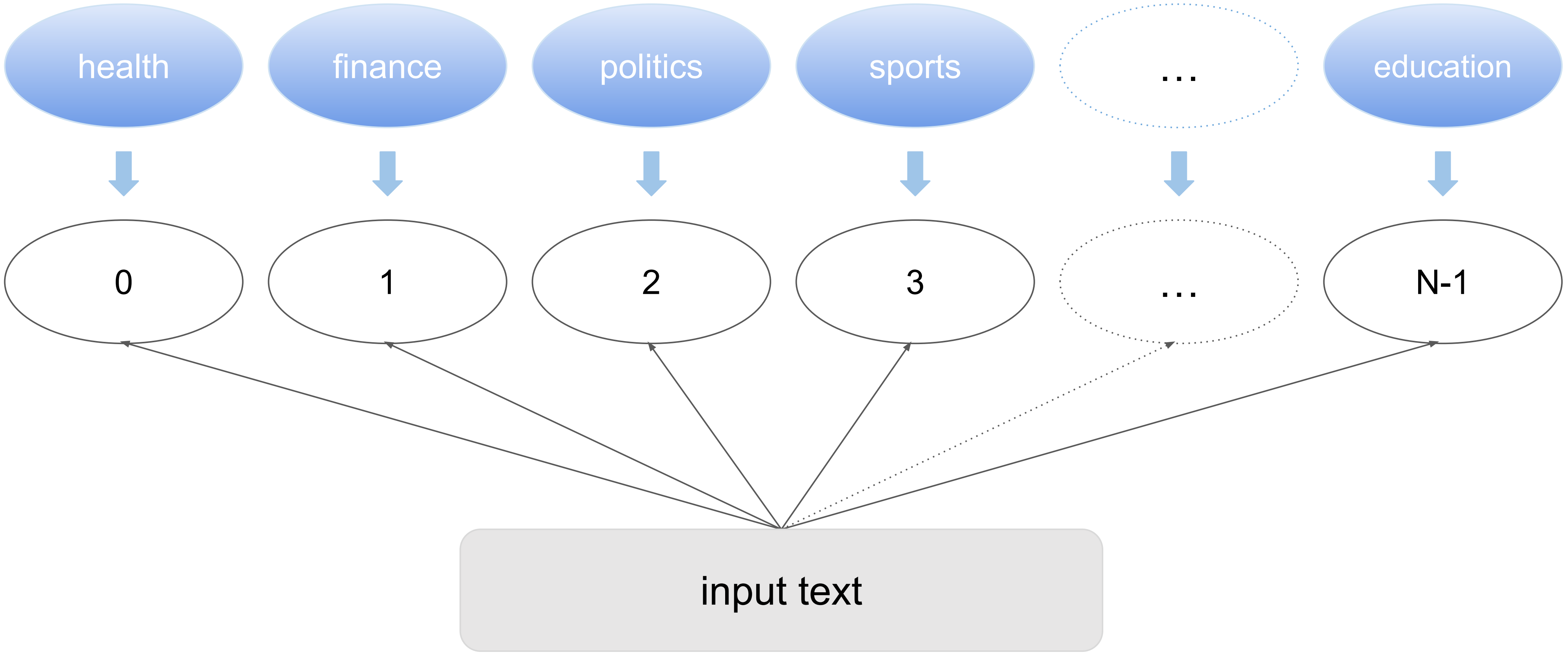
How do conventional classifiers work



How do conventional classifiers work

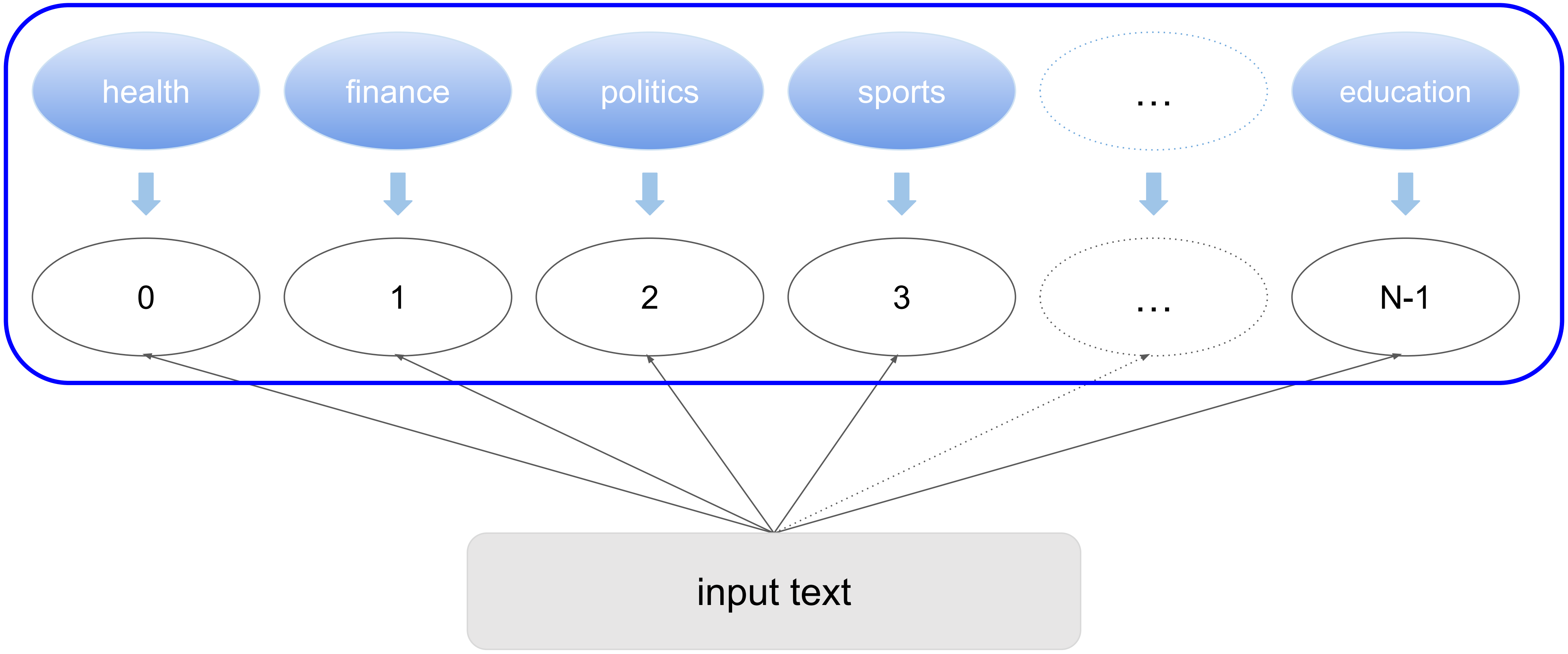


How do conventional classifiers work

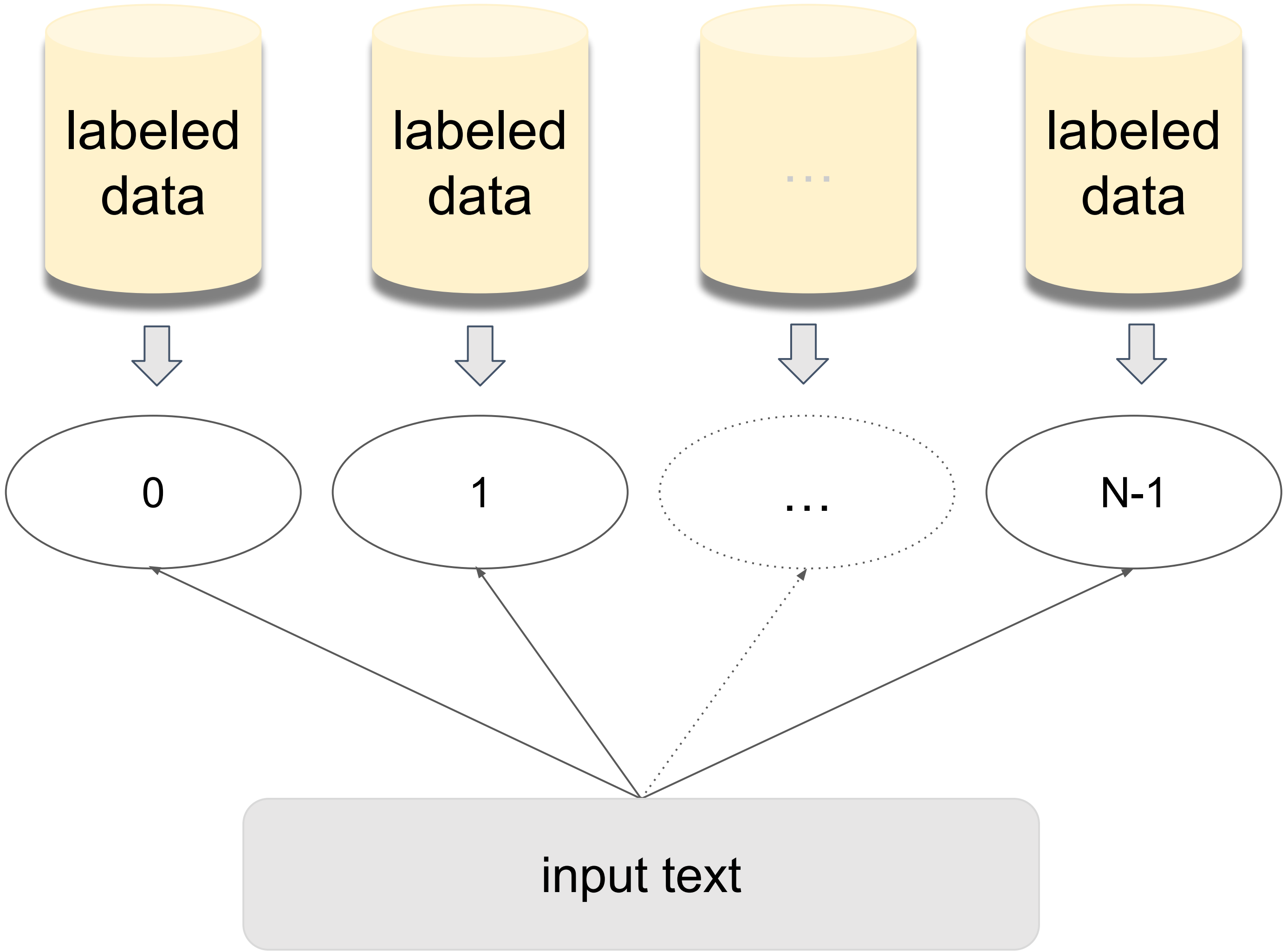


How do conventional classifiers work

real labels are converted into indices; label semantics are missing



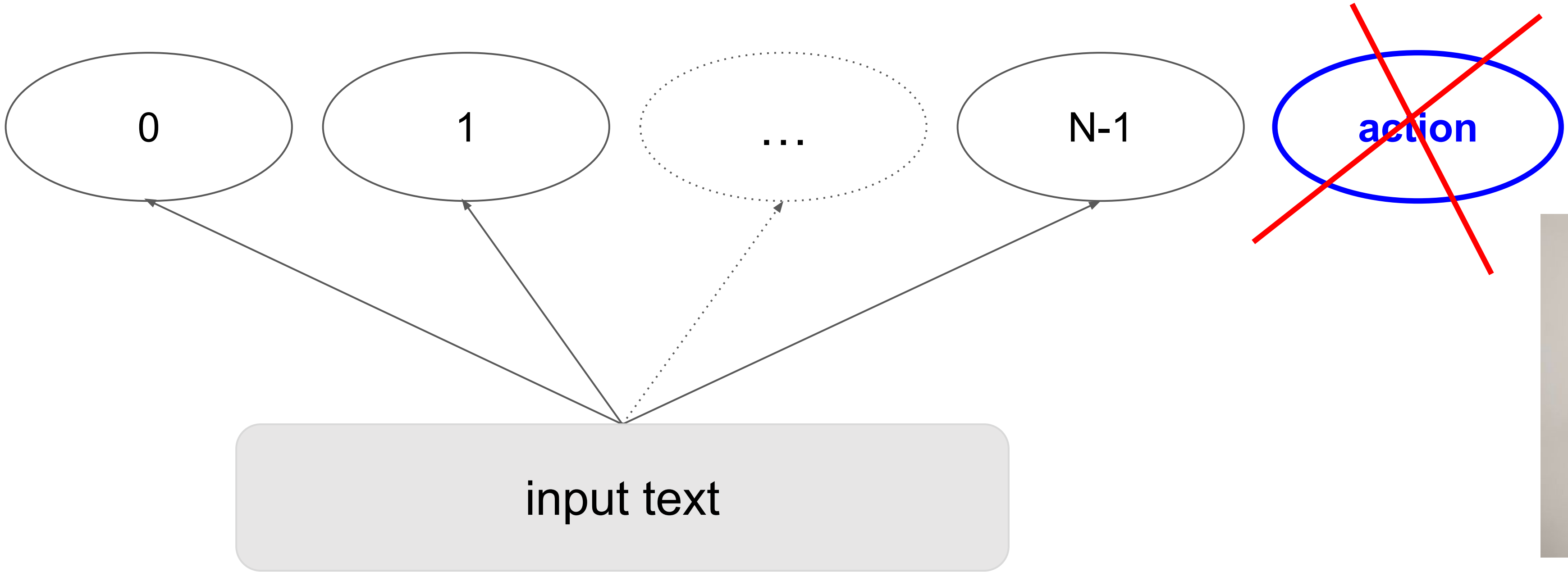
Issues of conventional classifiers



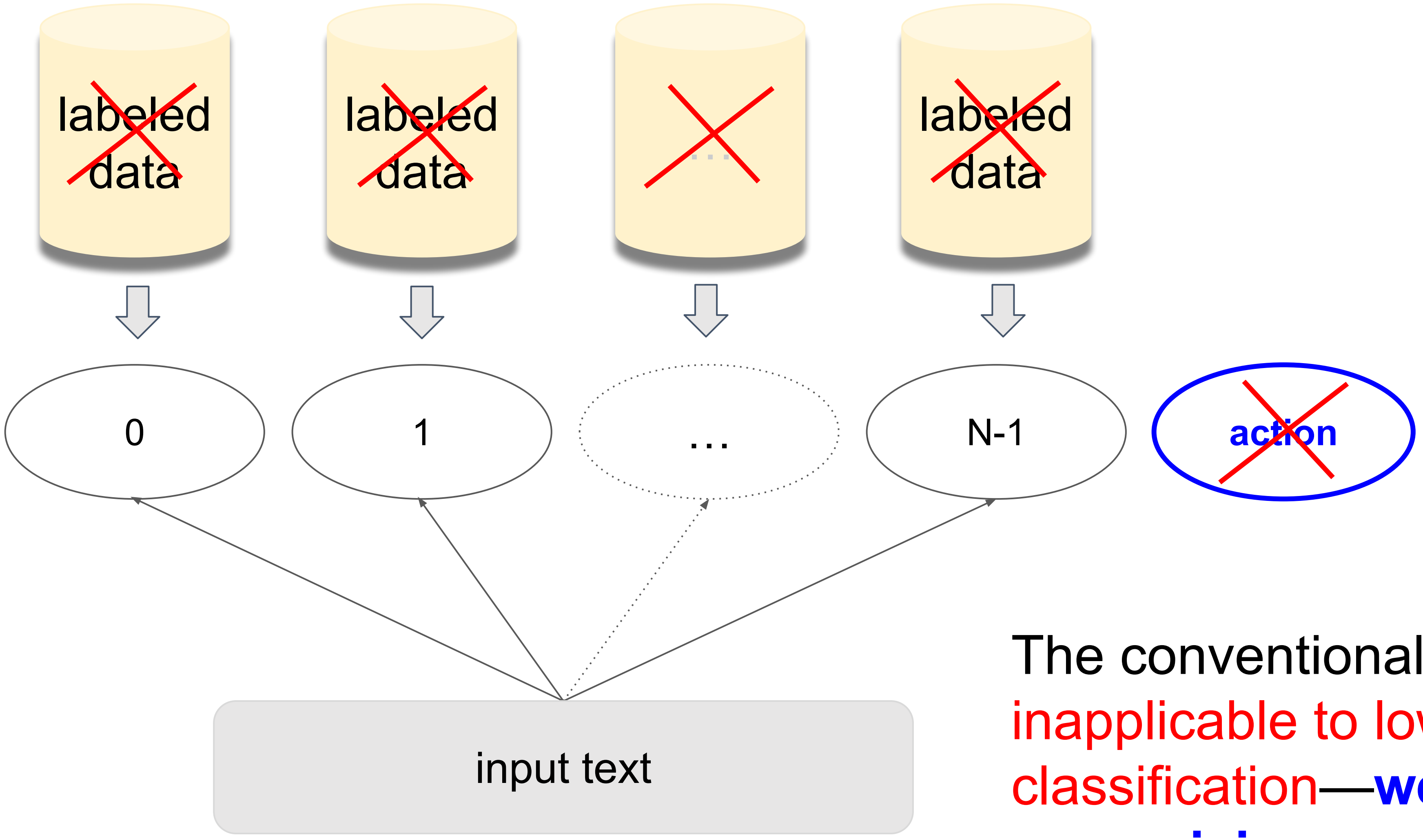
1) needs a large number of labeled examples to learn existing types

Issues of conventional classifiers

1) needs a large number of labeled examples to learn existing types
2) cannot generalize to new types



Issues of conventional classifiers



The conventional approach is inapplicable to low-annotation text classification—**we need indirect supervision**

NLI for zero-shot text classification

Zero-shot text classification

Natural language inference

0 (health)

1 (anger)

2 (accident)

3 (crime)

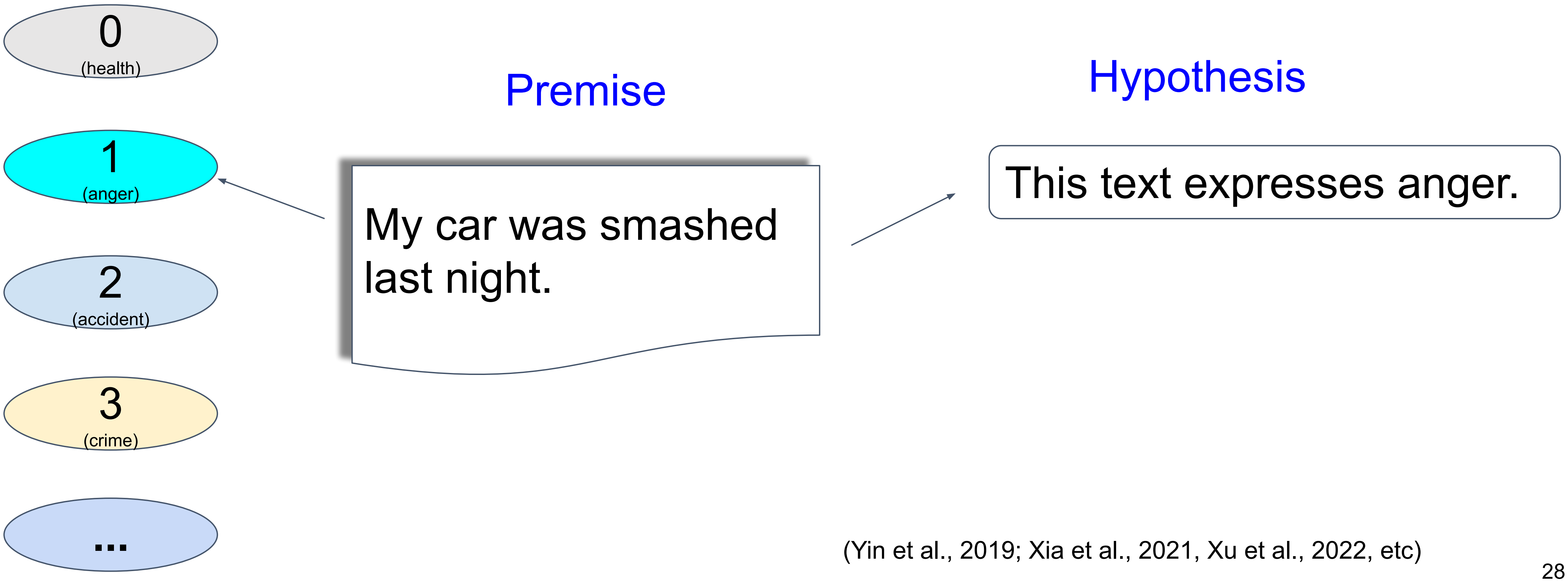
...

My car was smashed last night.

(Yin et al., 2019; Xia et al., 2021, Xu et al., 2022, etc)

NLI for zero-shot text classification

Zero-shot text classification | **Natural language inference**



(Yin et al., 2019; Xia et al., 2021, Xu et al., 2022, etc)

NLI for zero-shot text classification

Zero-shot text classification

Natural language inference

0
(health)

1
(anger)

2
(accident)

3
(crime)

...

Premise

My car was smashed last night.

Hypothesis

This text expresses anger.

This text is about accident.

(Yin et al., 2019; Xia et al., 2021, Xu et al., 2022, etc)

NLI for zero-shot text classification

Zero-shot text classification

0 (health)

1 (anger)

2 (accident)

3 (crime)

...

Natural language inference

Premise

My car was smashed last night.

Hypothesis

This text expresses anger.

This text is about accident.

We can keep the original label strings, or use other related words
— **label verbalizer** (Schick and Schütze, 2020)

NLI system—pretraining

NLI datasets:

- MNLI (Williams et al., 2018)
- ANLI (Nie et al., 2020)
- SNLI (Bowman et al., 2015)
- DocNLI (Yin et al., 2021)
- SciTail (Khot et al., 2018)
- RTE (Dagan et al. 2005)
- …

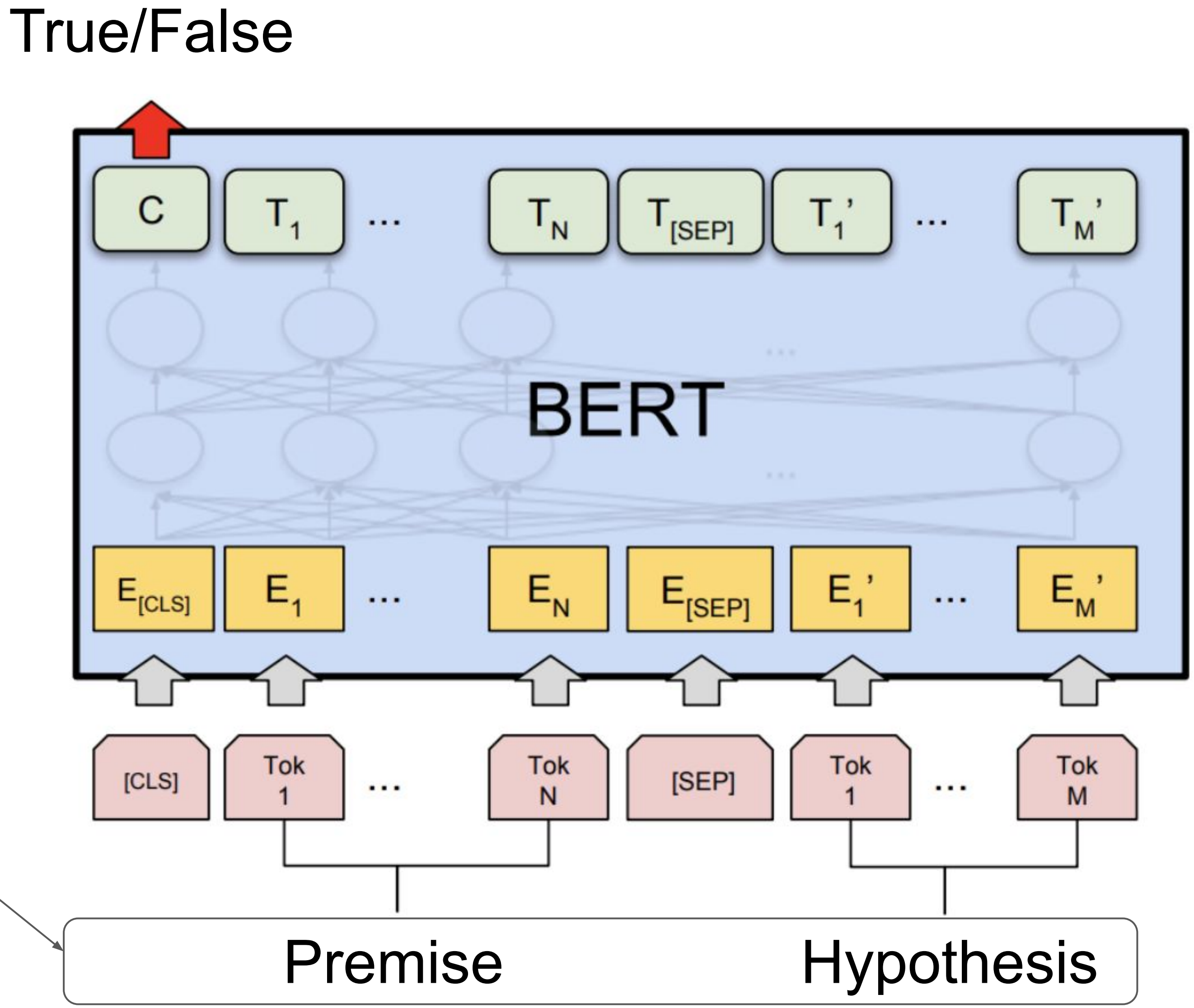
PLMs:

- BERT, RoBERTa, T5, etc.

Training pairs

- are from the NLI datasets;
- are not specific to any target labels

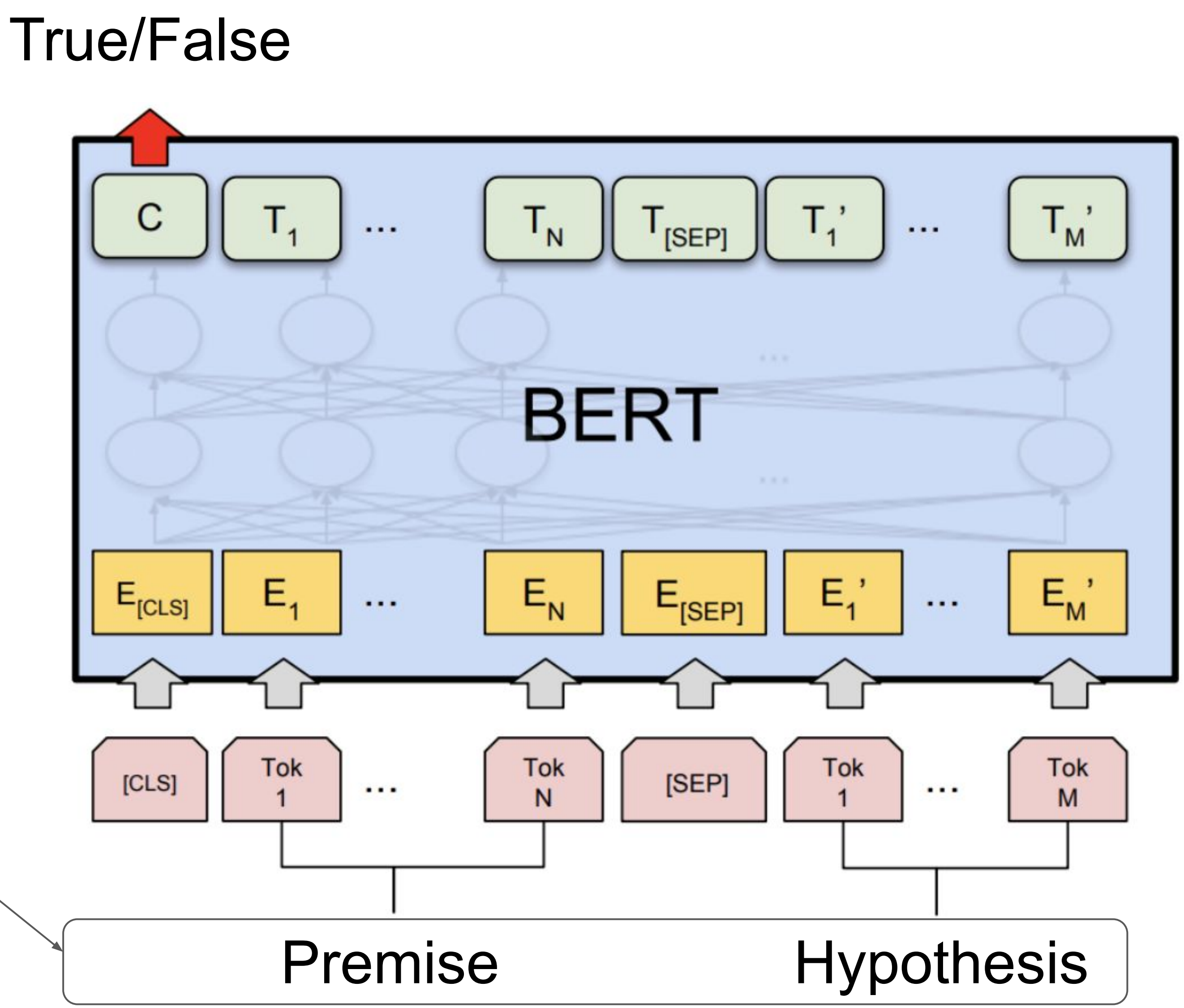
(Yin et al., 2019; Xia et al., 2021, Xu et al., 2022, etc)



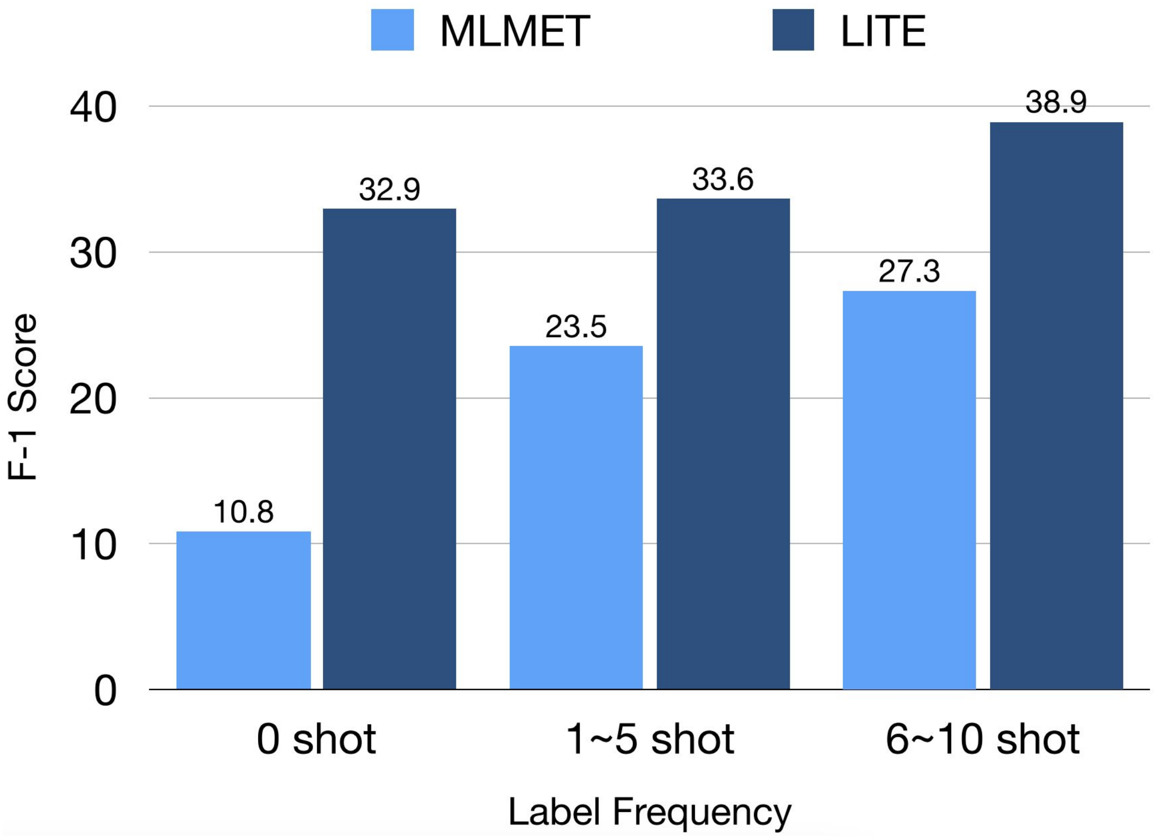
NLI system—finetuning (optional)

PLMs:
- ❑ pretrained NLI model

Training pairs
- ❑ **are from the target task** (if available);
  - topic classification (Yin et al., 2019)
  - multi-choice QA and corefernece (Yin et al., 2020)
  - intent identification (Xia et al., 2021)
  - stance detection (Xu et al., 2022)
  - ultra-fine entity typing (Li et al., 2022)
  - event argument extraction (Sainz et al., 2022)
  - (biomedical) relation extraction (Sainz et al., 2021; Xu et al., 2023)
  - …
- ❑ **are in the NLI format**

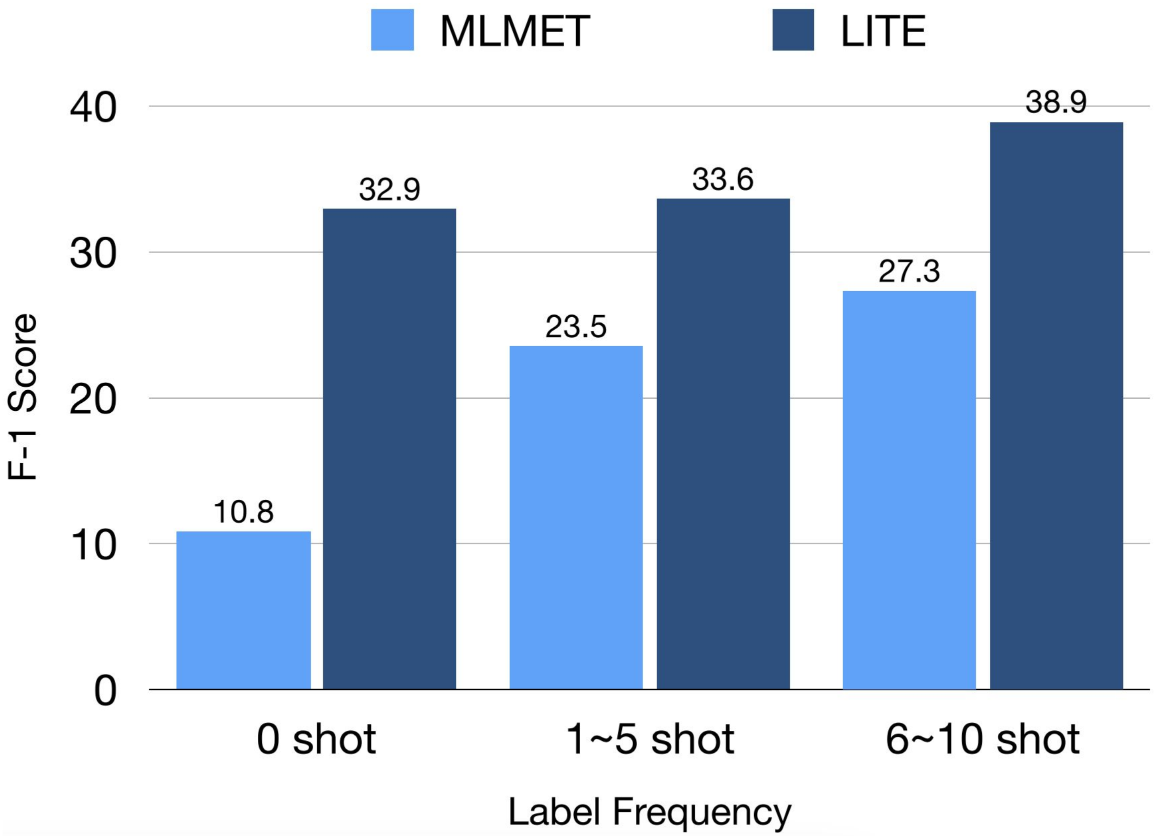


Performance

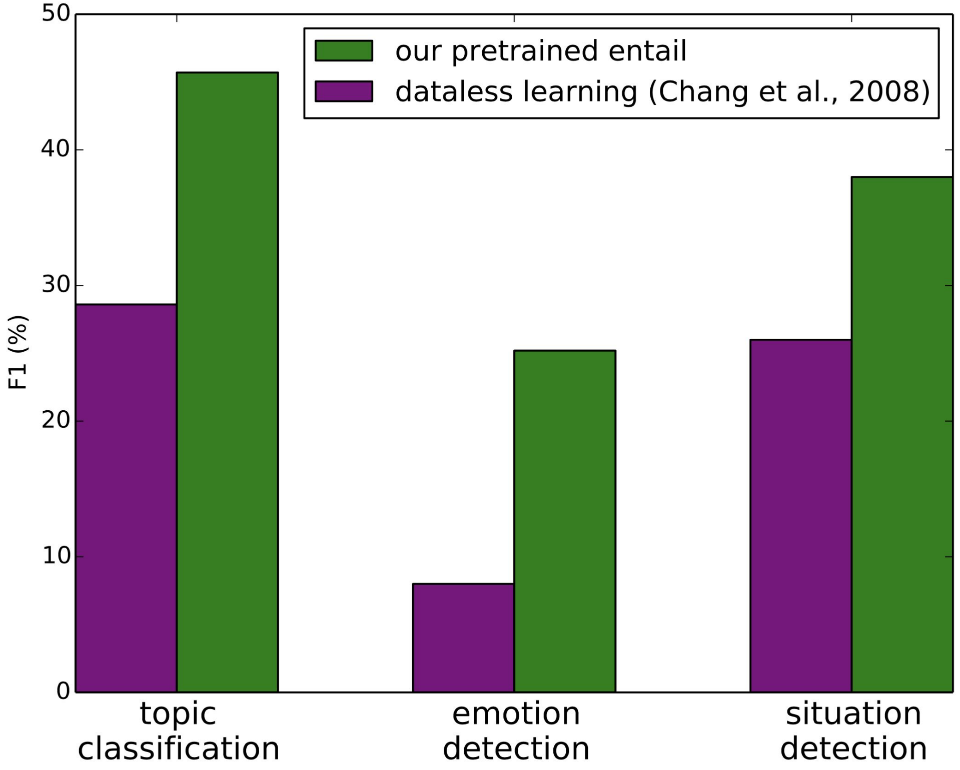


Ultra-fine entity typing ([Li et al., TACL 2022](#))
NLI can handle zero-shot & few-shot

Performance



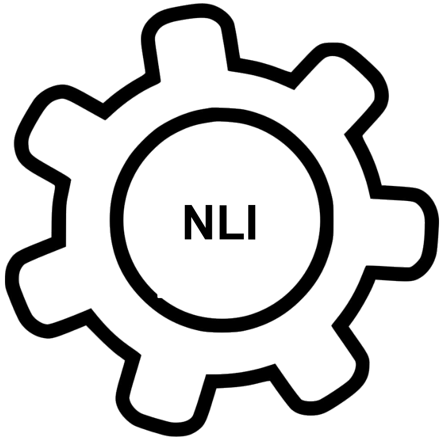
Ultra-fine entity typing (Li et al., TACL 2022)
NLI can handle zero-shot & few-shot



Zero-shot text classification (Yin et al., EMNLP 2019). **One NLI system can handle various zero-shot tasks**

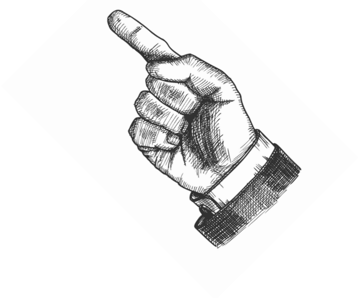
NLI-based indirect supervision: outline

- ❑ Implementation & Applications



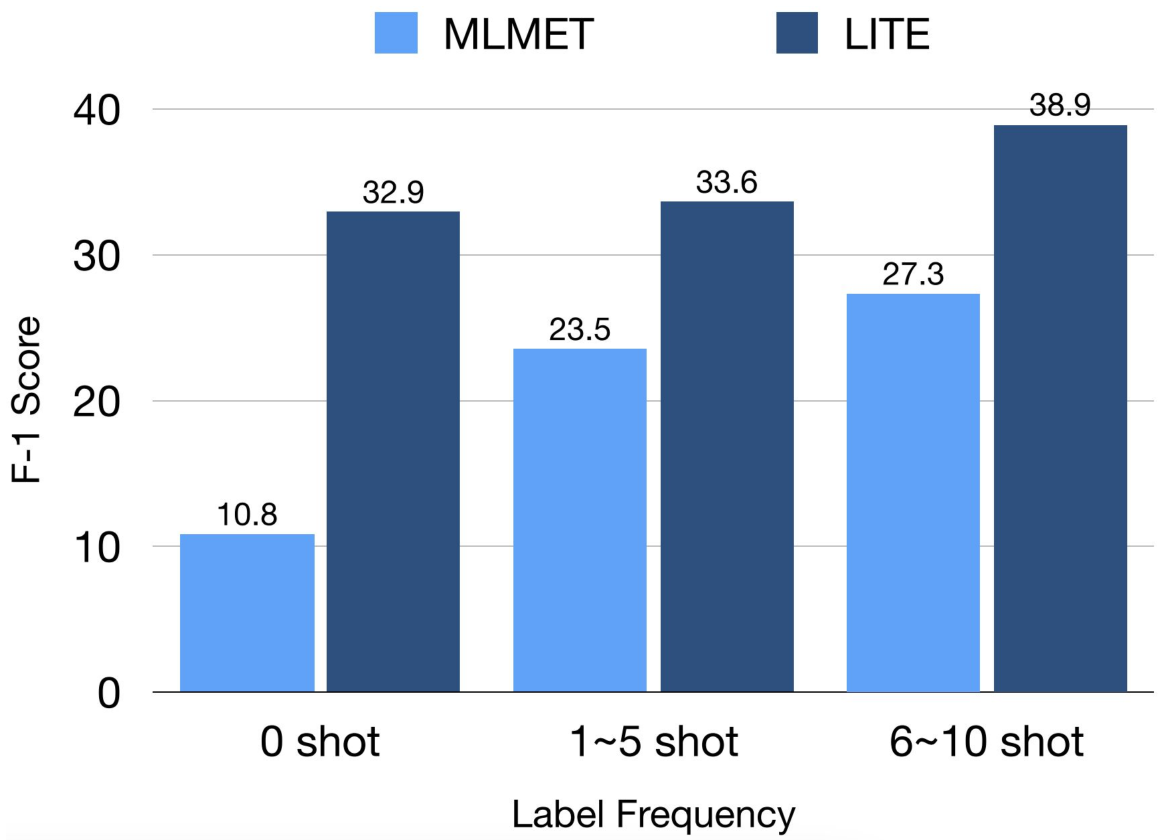
- ❑ Benefits

- ❑ Challenges & Solutions



Indirect supervision from NLI: benefits

- ❏ Reduce task-specific annotation requirements and address few-shot and zero-shot scenarios in a unified approach.



(Li et al., TACL 2022)

Indirect supervision from NLI: benefits

- ❑ Reduce task-specific annotation requirements and address few-shot and zero-shot scenarios in a unified approach.
- ❑ Facilitate cross-task transferability, encompassing not only NLI to target tasks but also task A to task B.

	original task	domain	premise length	hypothesis length
ANLI	NLI	various (wiki, news, etc.)	multi-sentence (20∼94 words)	single sentence (4∼18 words)
SQuAD	QA	wiki	paragraph (27∼237 words)	single sentence (6∼22 words)
DUC (2001)	summarization	news	doc. (124∼879 words)	multi-sent (80 ∼100 words)
CNN/Daily Mail	summarization	news	doc. (247∼652 words)	3∼4 sent. (40∼50 words)
Curation	summarization	news	doc. (229∼842 words)	multi-sent (64∼279 words)

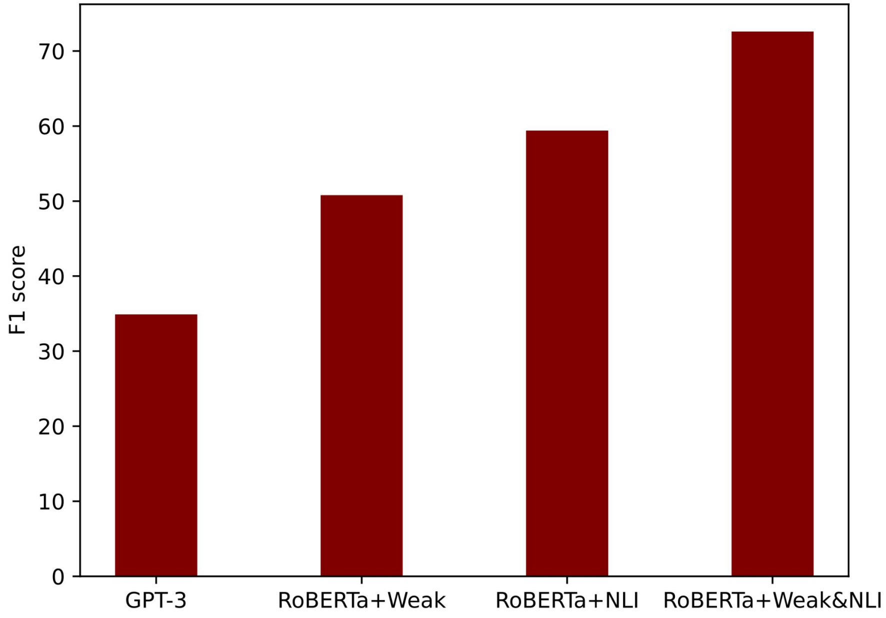
DocNLI (Yin et al., 2021) converted QA, summarization tasks as NLI-style source tasks

		FEVER	MCTest	
		binary	v160	v500
	random	50.00	25.00	25.00
pretrain	MNLI	86.64	75.41	70.66
	ANLI	87.51	82.50	78.66
	DocNLI	88.84	90.00	85.83
	+finetune	**89.44**	**90.83**	**90.66**
Prior state-of-the-art		–	80.00	75.50

DocNLI generalizes to distinct target tasks

Indirect supervision from NLI: benefits

- ❑ Reduce task-specific annotation requirements and address few-shot and zero-shot scenarios in a unified approach.
- ❑ Facilitate cross-task transferability, encompassing not only NLI to target tasks but also task A to task B.
- ❑ Enhance the feasibility and potential of employing smaller PLMs.

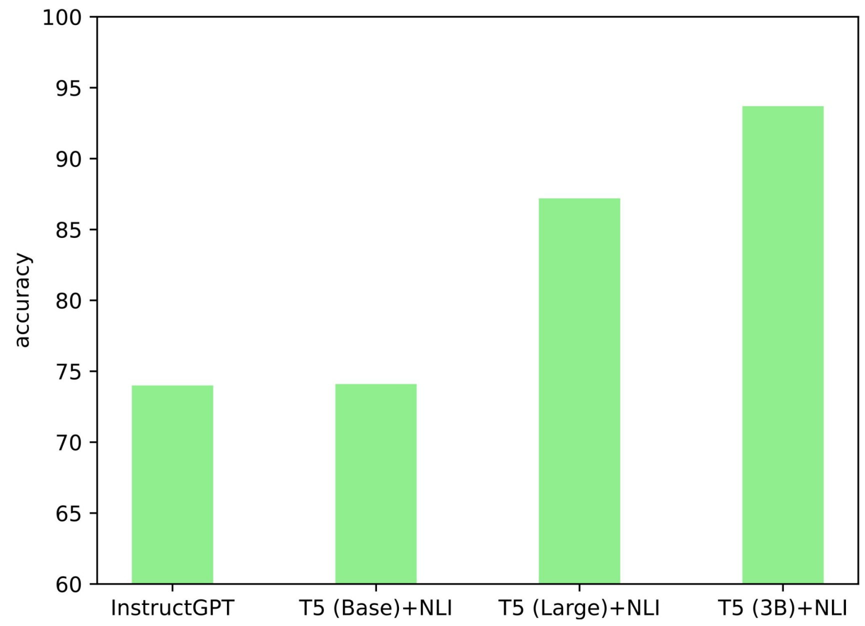


"*[OpenStance: Real-world Zero-shot Stance Detection](#)*" (Xu et al., CoNLL'22)

- ❑ Zero-shot
- ❑ RoBERTa (355M) + weak&NLI > GPT-3 (175B)

Indirect supervision from NLI: benefits

- ❑ Reduce task-specific annotation requirements and address few-shot and zero-shot scenarios in a unified approach.
- ❑ Facilitate cross-task transferability, encompassing not only NLI to target tasks but also task A to task B.
- ❑ Enhance the feasibility and potential of employing smaller PLMs.

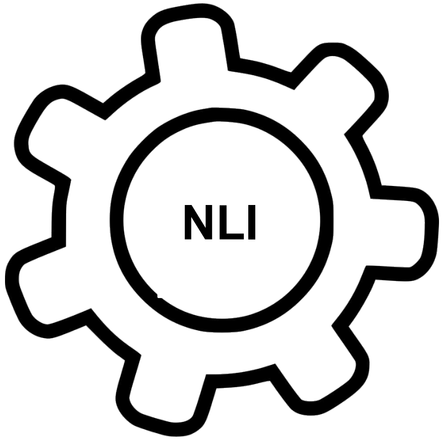


"*ZARA: Improving Few-Shot Self-Rationalization for Small Language Models*" (Chen et al., 2023)

- ❑ Few-shot
- ❑ NLI helps automatic data generation
- ❑ T5 (base)+NLI (2M) ≈ InstructGPT (175B)
- ❑ T5 (large)+NLI (7M) > InstructGPT (175B)
- ❑ T5 (3B)+NLI (2.7B) > InstructGPT (175B)

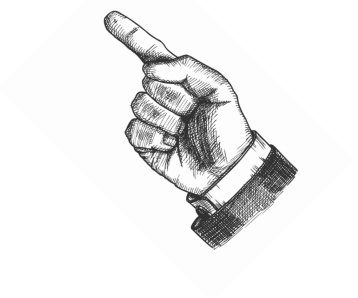
NLI-based indirect supervision: outline

❑ Implementation & Applications



❑ Benefits

❑ Challenges & Solutions



Challenge #1: domain discrepancy

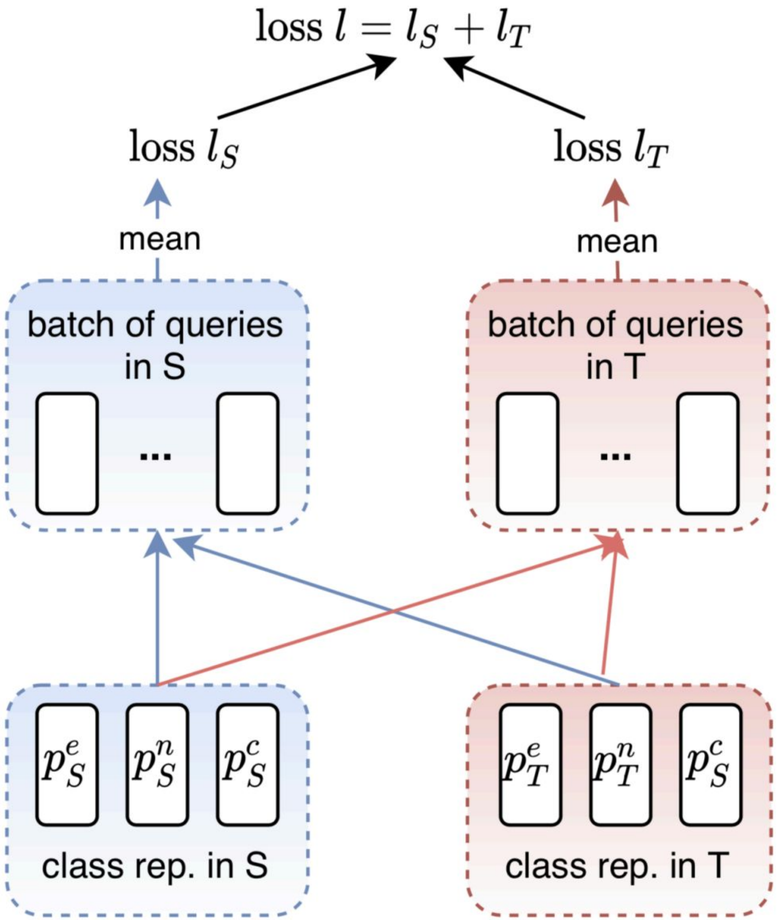
Domains of the target problems (T) often differ from that of NLI datasets (S)

Traditional solution: pretrain on S +finetune on T (i.e., STILTS (Phang et al., 2018))

Challenge #1: domain discrepancy

Domains of the target problems (T) often differ from that of NLI datasets (S)

Traditional solution: pretrain on S +finetune on T (i.e., STILTS (Phang et al., 2018))



Solution I:

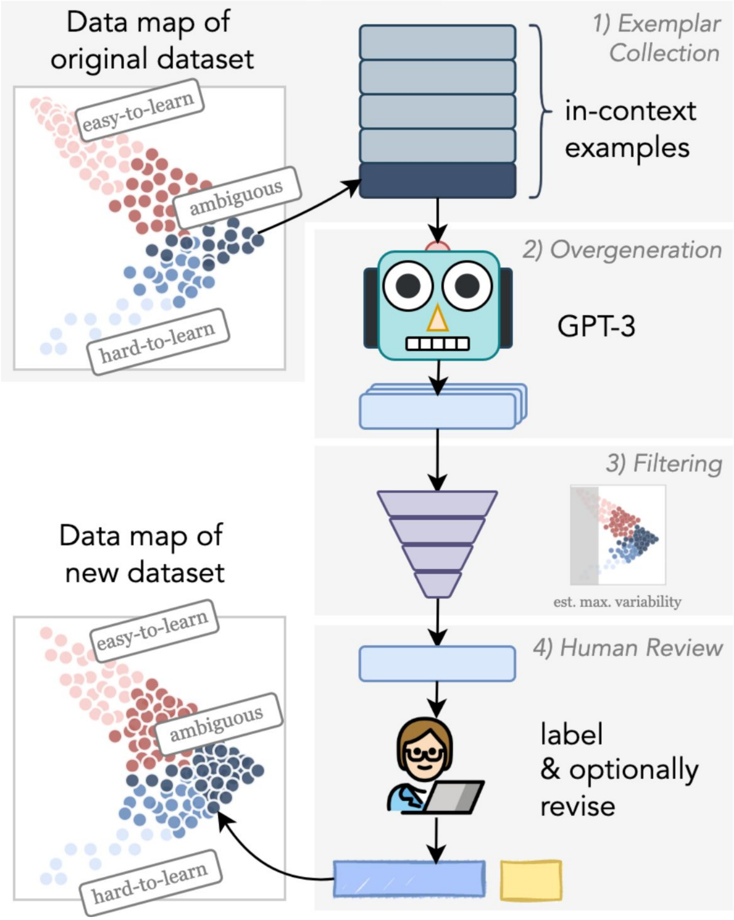
- T is few-shot → imitate few-shot learning on S (meta-learning)
- Novel strategy: predictions on T (or S) depend on the signals of both T and S
- A solution from the algorithm perspective

Universal Natural Language Processing with Limited Annotations: Try Few-shot Textual Entailment as a Start (Yin et al., EMNLP 2020)

Challenge #1: domain discrepancy

Domains of the target problems (T) often differ from that of NLI datasets (S)

Traditional solution: pretrain on S +finetune on T (i.e., STILTS (Phang et al., 2018))



Solution II:

- ❑ Human annotations: correctness, diversity
- ❑ PLMs: creative writing
- ❑ Use GPT-3 to generate new examples for reasoning patterns that are challenging
- ❑ A solution from the data perspective

WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation
(Liu et al., Findings of EMNLP 2022)

Challenge #2: inefficiency in testing

Each input needs to infer a large number of output-specific hypotheses

- ❏ Ultra-fine entity typing (Choi et al. 2018): 10K labels take the NLI model (Li et al., 2022) 35 seconds for each test instance and about 19.4 hours to infer the entire test set.

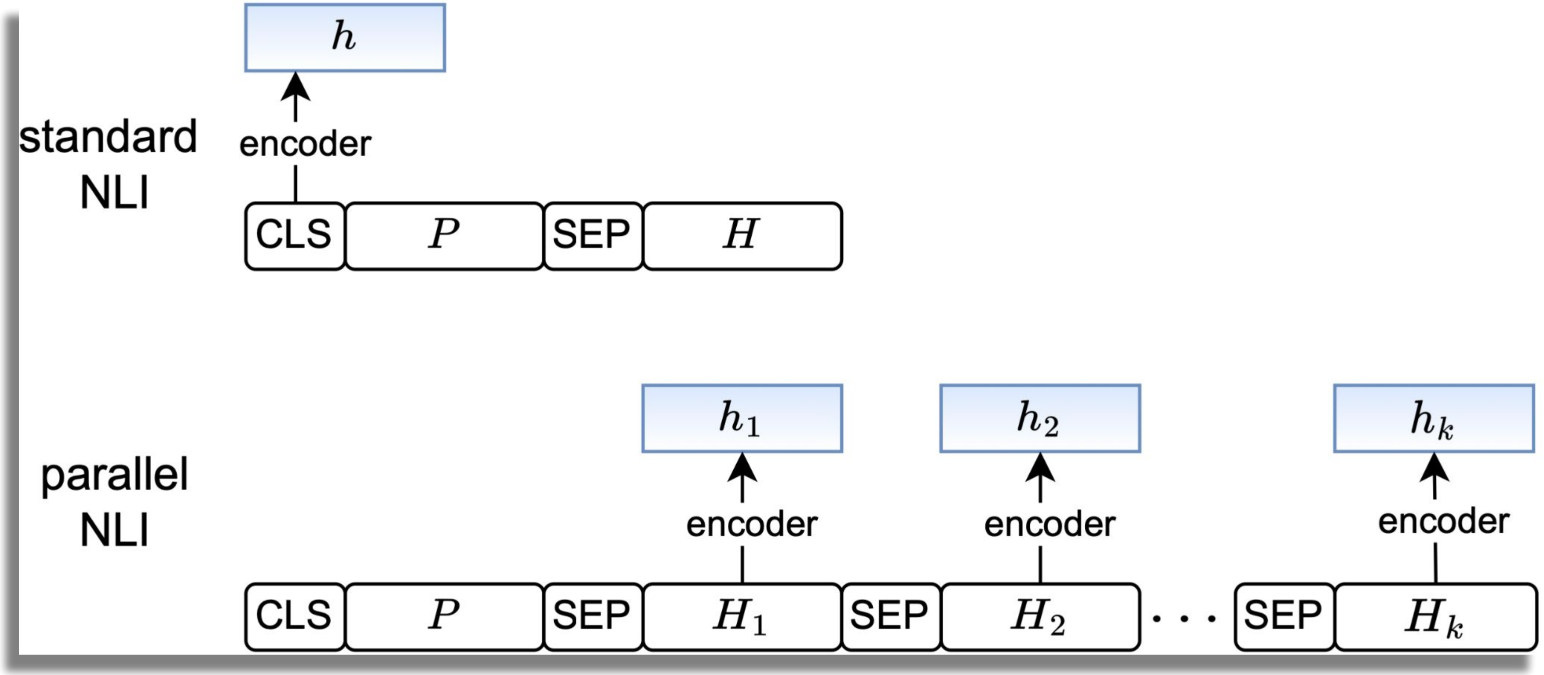
Challenge #2: inefficiency in testing

Each input needs to infer a large number of output-specific hypotheses

- ❑ Ultra-fine entity typing (Choi et al. 2018): 10K labels take the NLI model (Li et al., 2022) 35 seconds for each test instance and about 19.4 hours to infer the entire test set.

Solution:

- ❑ Pairwise inference → group-wise inference
- ❑ Assumption: hypotheses exhibit a binary polarity irrespective of their competitors.



Learning to Select from Multiple Options (Du et al., AAAI'23)
Recall, Expand and Multi-Candidate Cross-Encode: Fast and Accurate Ultra-Fine Entity Typing (Jiang et al., Arxiv, 2022)

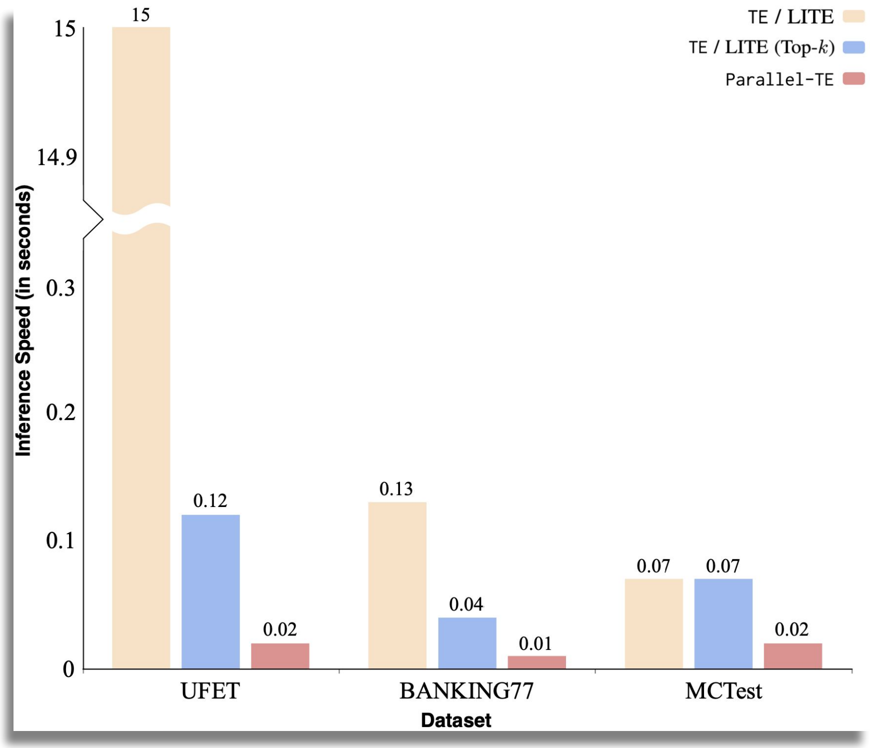
Challenge #2: inefficiency in testing

Each input needs to infer a large number of output-specific hypotheses

- Ultra-fine entity typing (Choi et al. 2018): 10K labels take the NLI model (Li et al., 2022) 35 seconds for each test instance and about 19.4 hours to infer the entire test set.

Solution:

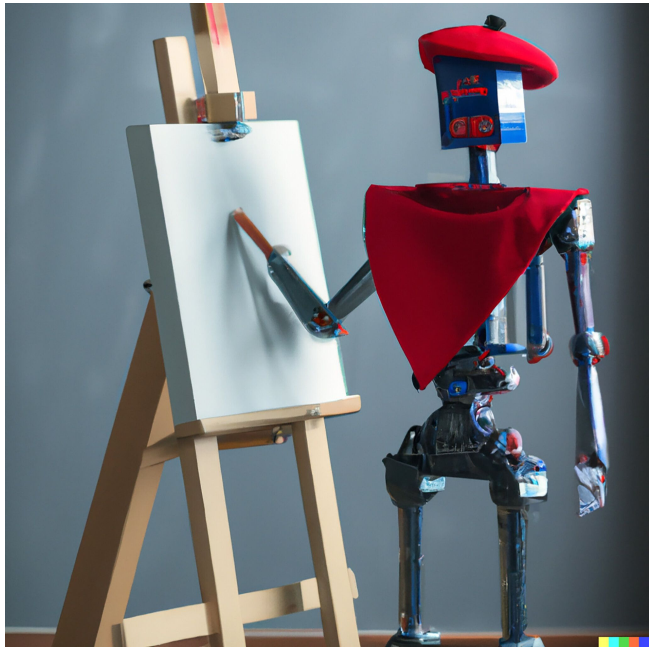
- Pairwise inference → group-wise inference
- 15 seconds → 0.02 seconds (per example)



[Learning to Select from Multiple Options](#) (Du et al., AAAI'23)
[Recall, Expand and Multi-Candidate Cross-Encode: Fast and Accurate Ultra-Fine Entity Typing](#) (Jiang et al., Arxiv, 2022)

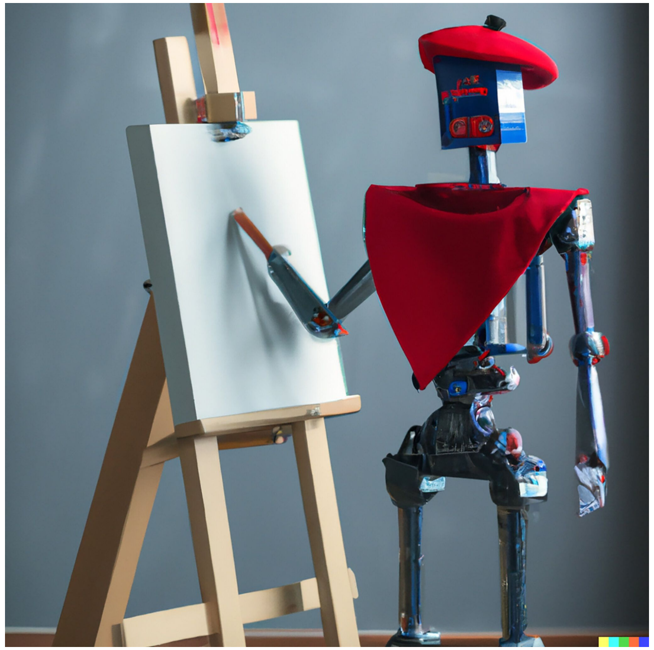
Challenge #3: cannot discover new labels

- We often pre-define the label set for classification tasks
- At times, we may want the model to generate some new labels for the input to "surprise" us



Challenge #3: cannot discover new labels

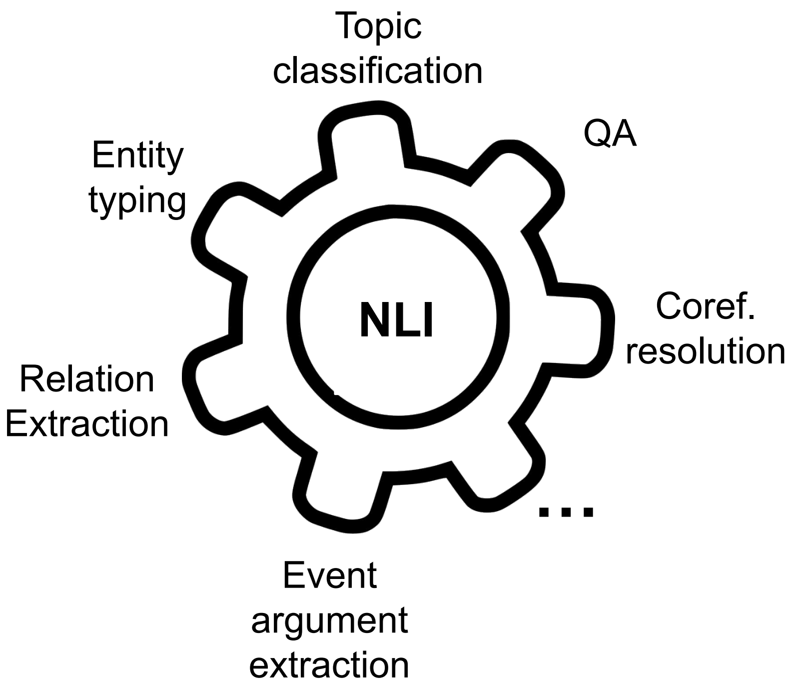
- We often pre-define the label set for classification tasks
- At times, we may want the model to generate some new labels for the input to "surprise" us



Will be addressed by the next section of our tutorial

Recap of indirect supervision from NLI

Implementation & Applications



Benefits

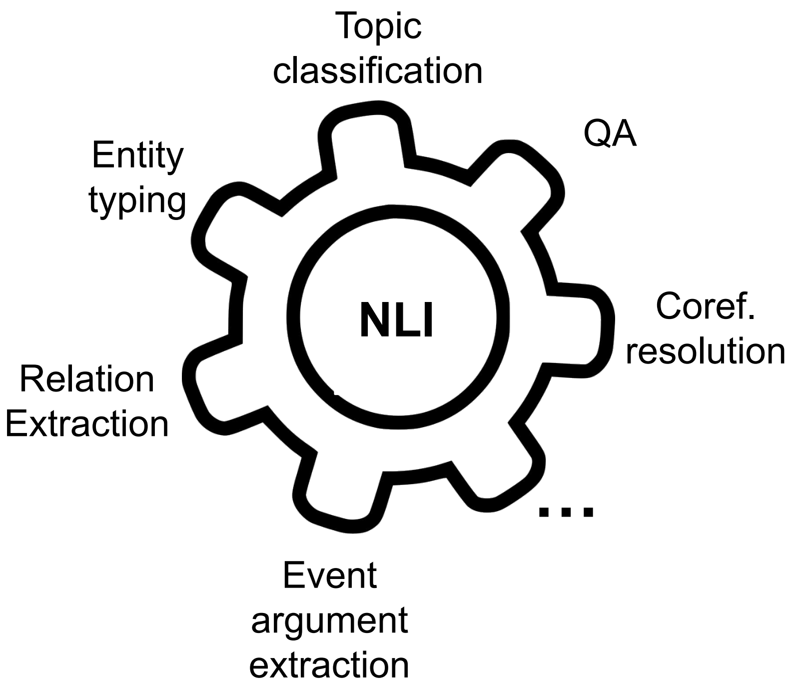
- ❑ Scarce-annotation NLP
- ❑ Cross-task transferability
- ❑ Maximize the potential of small PLMs

Challenges & Solutions

- ❑ Domain discrepancy (solutions by algorithm and data threads)
- ❑ Inefficiency in testing (parallel-NLI)
- ❑ cannot discover new labels (next chapter…)

Recap of indirect supervision from NLI

Implementation & Applications



Benefits

- ❑ Scarce-annotation NLP
- ❑ Cross-task transferability
- ❑ Maximize the potential of small PLMs

Challenges & Solutions

- ❑ Domain discrepancy (solutions by algorithm and data threads)
- ❑ Inefficiency in testing (parallel-NLI)
- ❑ cannot discover new labels (next chapter…)

Thank You