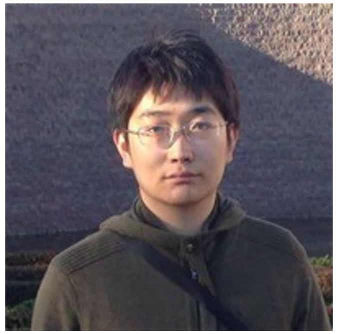
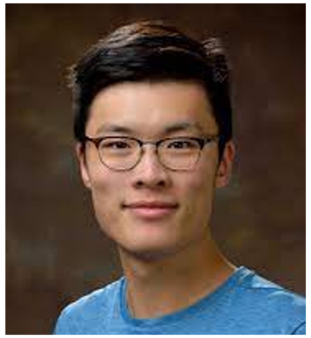
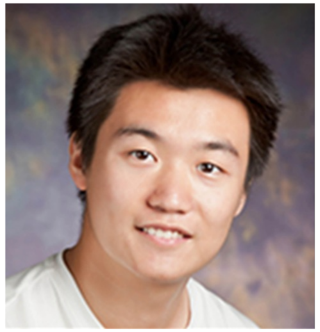
PennState College of Engineering | USC Viterbi School of Engineering | amazon | UCLA Samueli School of Engineering | Penn

Indirectly Supervised Natural Language Processing


Wenpeng Yin


Muhao Chen


Ben Zhou


Qiang Ning


Kai-Wei Chang


Dan Roth

July 9, 2023

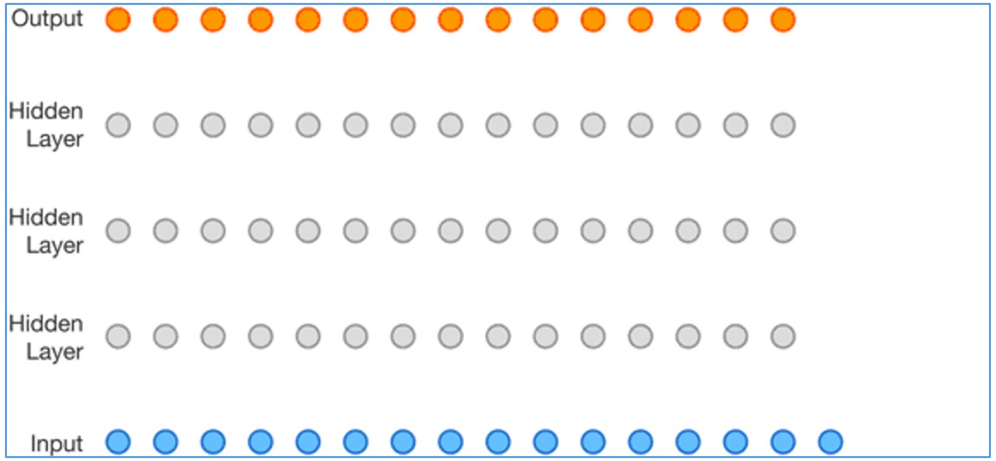
ACL Tutorials

Indirectly Supervised Natural Language Processing

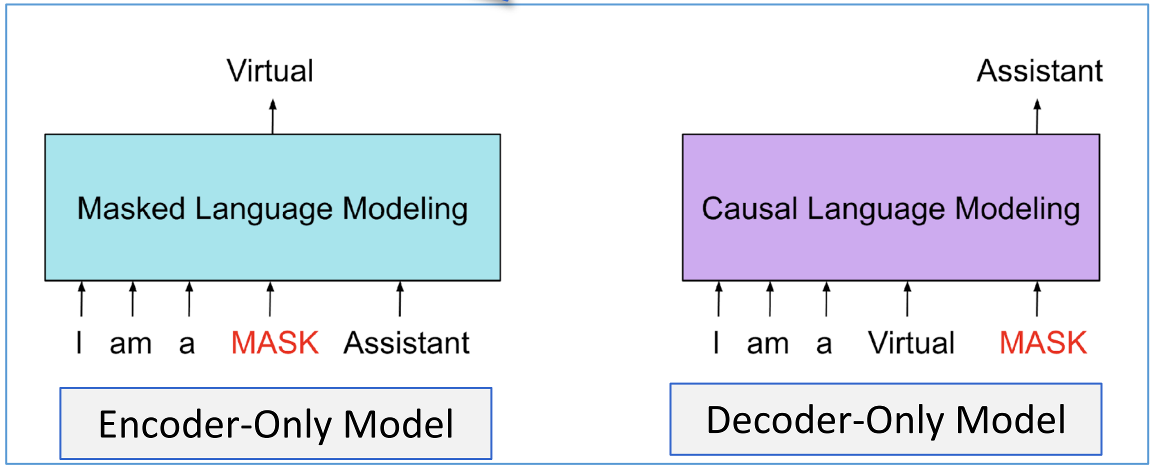
61 ACL 2023

Do We Need Supervision?

- All parameters in NLP networks are initialized via **pretraining**.
- Pretraining methods hide parts of the input and train the model to reconstruct those parts.



This **self-supervision** is the key reason for the original excitement with LLMs: the hope that we can get around the need to annotate a lot of data for supervised machine learning.



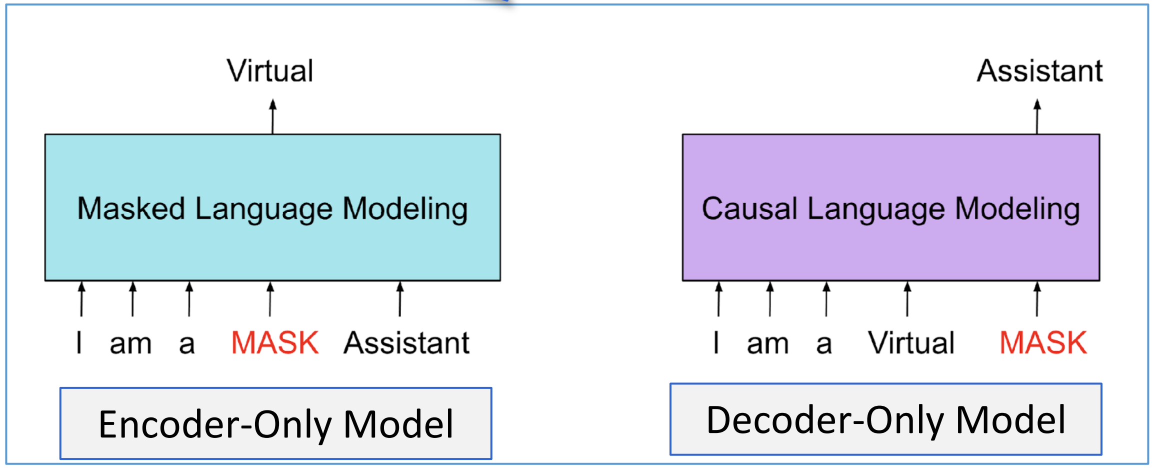
But this is **no longer true**.
All the very large models use huge amounts of **supervision** and **RLHF** (Reinforcement Learning with Human Feedback) data that is more costly than earlier supervision protocols.

What Have We Learned?

- We learned that it's possible to generate effective supervision
  - **Without** manual annotation
  - **Without** training directly for the task at hand
- Not new
  - All the literature on Context Sensitive Spelling (accept/except; the/that,...) and (ESL) Text Correction is based on this self-supervision paradigm [Golding & Roth ICML'96]
- **Needs** to be **generalized**
  - And applied to many machine learning tasks
  - Since "simplistic" self-supervision isn't good enough to support most tasks

This tutorial

This **self-supervision** is the key reason for the original excitement with LLMs: the hope that we can get around the need to annotate a lot of data for supervised machine learning.



Incidental Supervision

- Data provides hints
  - Data exists independent of the task(s) at hand.
  - These “hints” are often sufficient to infer supervision signals for a range of tasks

Data Provides Hints

O'Hare must be in Chicago

- Feb 5 2017 Dozens of passengers heading to Chicago had to undergo additional screenings… It took Hesam Aamyab two tries to make it back to the United States from Iran. He is an Iranian citizen with a US visa who is doing post-doctoral research at UIC. …"Right now, I am in the USA and I'm very happy," Aamyab said. But now, he can't go back to Iran or anywhere else without risk. Other travelers shared the same worry. Asem Aleisawi was at O'Hare on Sunday to meet his wife who was coming in from Jordan.

    Learning/Supervision requires some level of reasoning to infer these weak signals

- Images

    Difficult to label objects/faces

    Easy to learn same/different

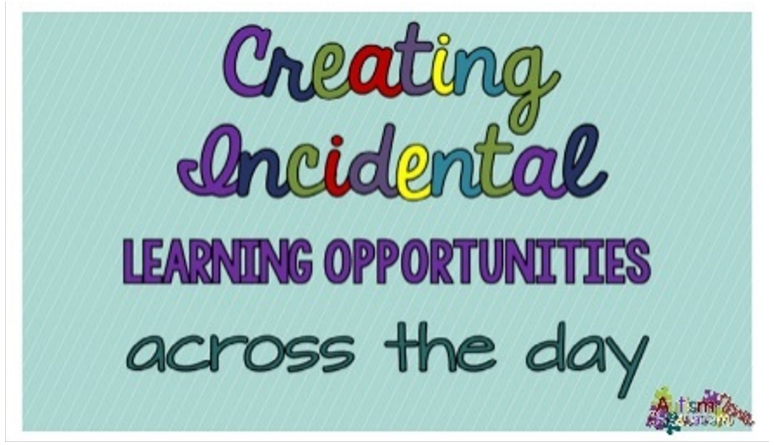
    - Two objects in the same image are different
    - Two consecutive video frames are likely to contain the same objects.

    A way to train intermediate representations that make the eventual labeling/prediction easier

    - Not the only way; but all realistic/scalable ways need to go through indirect supervision

Incidental Supervision [Roth, AAAI'17]

- Data provides hints
  - Data exists independent of the task(s) at hand.
  - These "hints" are often sufficient to infer supervision signals for a range of tasks
- Utilizing these hints
  - Can be substantially less costly than producing explicit annotation
  - More realistic – provides signals for tasks we haven't defined
  - Weak signals can be aggregated to produce higher quality signals
- Examples:
  - Classification Tasks
  - Structured Prediction Tasks
  - Multimodal Tasks
  - Commonsense
  - …..

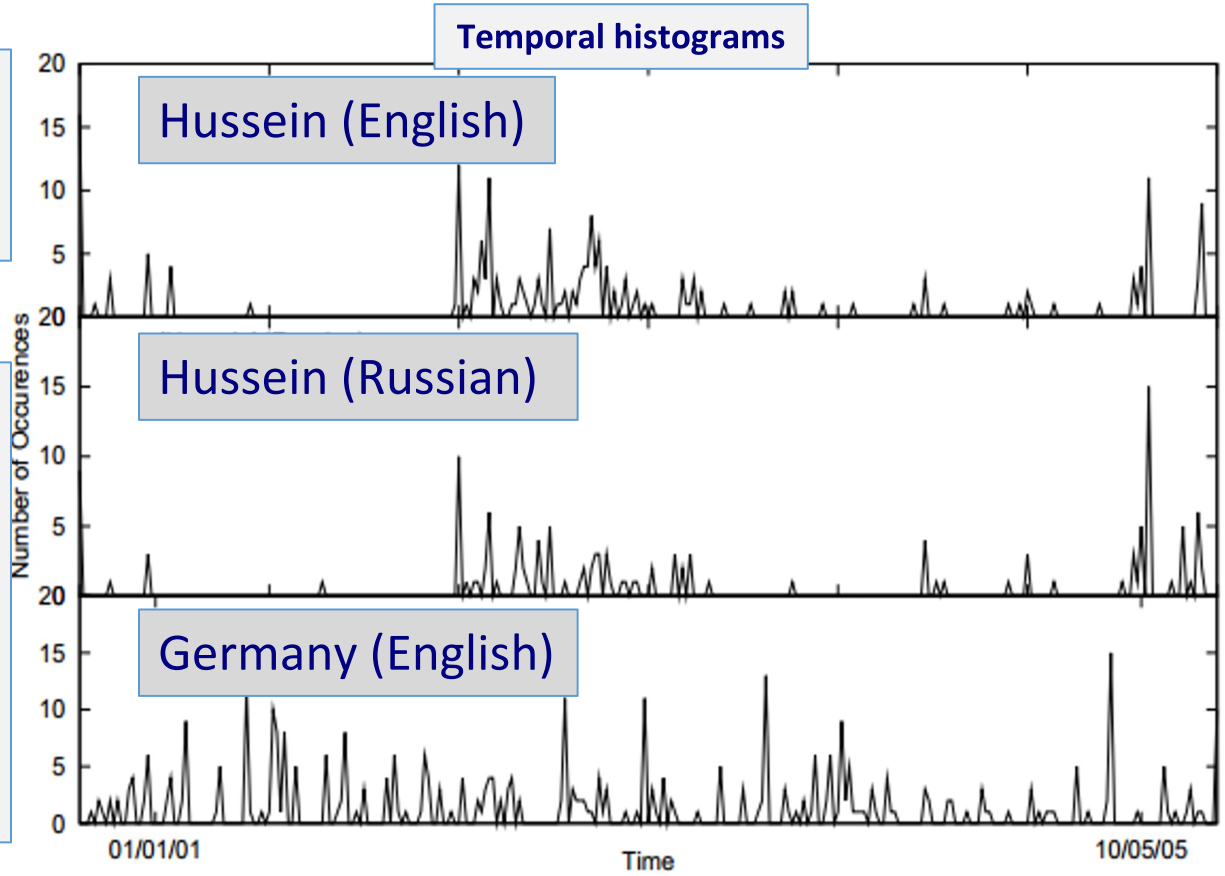


Diverse Supervision Signals

- Searching for supervision signals could be challenging.
- It is incidental, in the sense that it provides some signals that might be **co-related** with the target task and may not be useful all the time.

Assume a **comparable**, weakly temporally aligned news feeds.

Weak synchronicity provides a cue about the relatedness of (some) NEs across the languages, and can be exploited to associate them [Klementiev & Roth, 06,08]

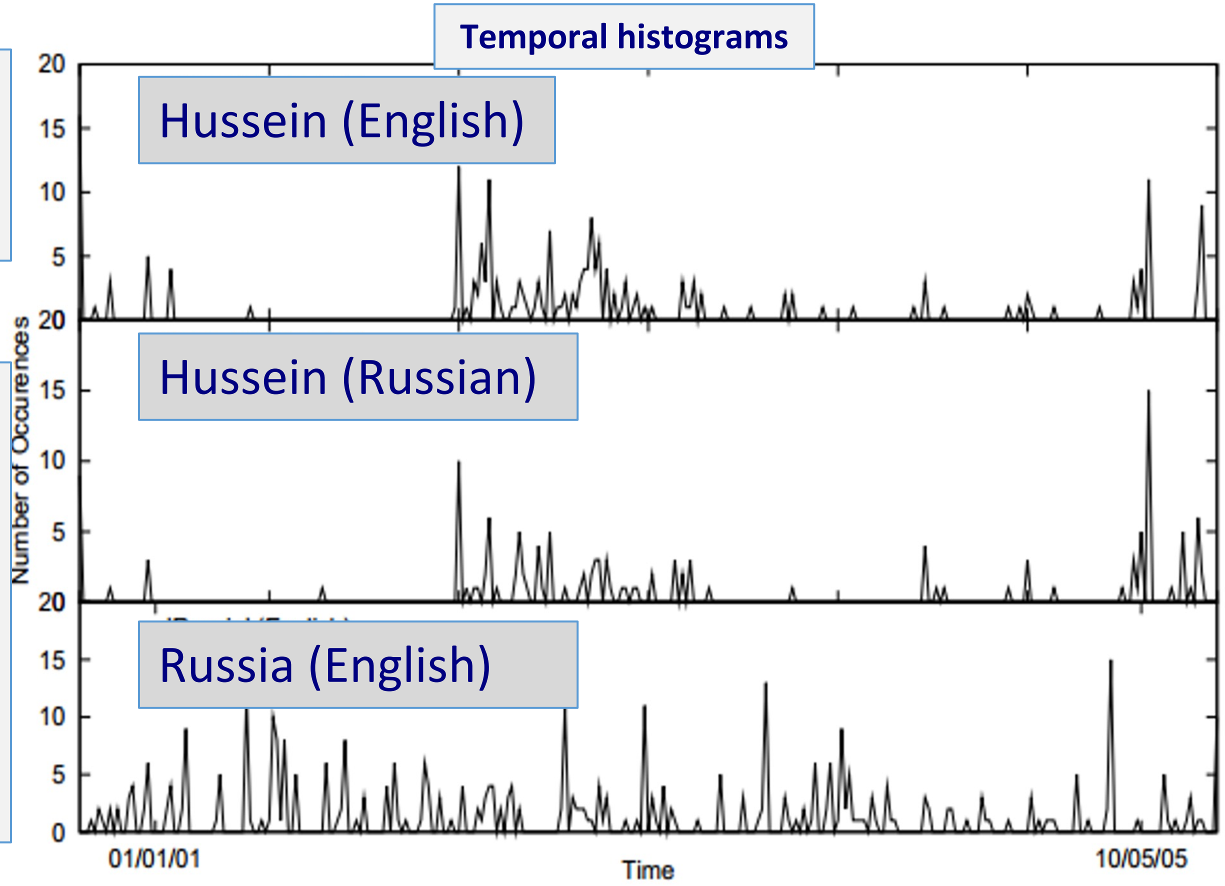


Diverse Supervision Signals

- By itself, this temporal signal may not be sufficient to support learning robust models.
- Along with weak phonetic signals, context, topics, etc. it can be used to get robust models.

Assume a **comparable**, weakly temporally aligned news feeds.

Weak synchronicity provides a cue about the relatedness of (some) NEs across the languages, and can be exploited to associate them [Klementiev & Roth, 06,08]



Solving NLP

We need many forms of indirect supervision.
We call this "**incidental supervision**" [Roth, AAAI'2017] since, unlike "standard supervision", we don't annotate for it, and we don't assume it is exhaustive; sometimes it's there and we use it.

- Even if we think/hope that generative AI will move us forward
- We still need supervision
  - Alignment with human expectations
  - Fine-tuning of models
  - Verification: is this piece of text supported by this evidence?
  - Classification into organization-specific taxonomies (e.g., medical)
  - Naming visual events
  - Text, images, video retrieval
  - ….
- We will never have enough annotated data to train all the models, for all the tasks we need
  - **We** do not learn by "training" on many examples
- Direct supervision is not scalable and, often, makes no sense
  - Complex tasks annotation is often impossible.

"understanding" Events

When and Where?

Fu et al. ACL'22

The language-world mapping problem

[BabySRL: Connor, Fisher & Roth, 2012]

Behavioral feedback is needed!

- Take inspiration from language acquisition?
  - Clearly, a lot of incidental supervision
  - Harder problems (understanding verbs) bootstrap from easier

“the world”

Smur! Rivvo della frowler.

“the language”

Scene 1

Topid rivvo den marplox.

Nouns identified

Blert dor marplox, arno.

[Topid rivvo den marplox.]

Scene 3

Marplox dorinda blicket.

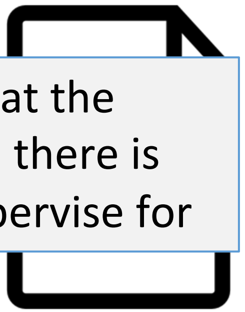
Scene n

Cross-situational observations

Think about this as comparable text

Sources of Incidental Supervision

- Representation-driven:
  - Label-aware: the basis of zero-shot **[Chang et al. AAAI'08]**
- Knowledge-driven
  - Enrichment of the text with existing knowledge
- Constraints-driven:
  - Expectation from the output
- Alignment-Driven:
  - comparable text; multimodal
- Behavior-driven:
  - It's end-to-end



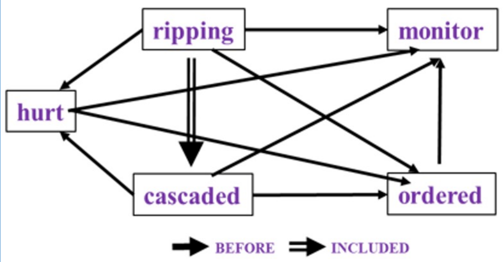
This document is about Tennis

It expresses joy

It constitutes harassment

If we know what the label "means", there is no need to supervise for it

In Los Angeles that lesson was brought home today when tons of earth **cascaded** down a hillside, **ripping** two houses from their foundations. No one was **hurt**, but firefighters **ordered** the evacuation of nearby homes and said they'll **monitor** the shifting ground until March 23rd.



We have **strong expectations** from the output:
(1) Transitivity (2) Expertations on "typical" order of events.

Who are the Europeans female tennis players who made the most money in the last 10 years?

~~SQL~~

1. Halep
2. Wozniacki
3. Azarenka
4. Kvitova
5. Kerber
6. …

This Tutorial

Tutorial Outline

■ Introduction Dan Roth	20 min.
■ Indirect Supervision from text classification Wenpeng Yin	30 + 5 min.
■ Indirect Supervision from text generation Muhao Chen	30 + 5 min.
■ Break	30 min.
■ Incidental Supervision from Natural Text Ben Zhou	25 + 5 min.
■ Theoretical Analysis of Incidental Supervision Qiang Ning	25 + 5 min.
■ Indirect Supervision from Multi-modalities Kai-Wei Chang	25 + 5 min.
■ Conclusion and Future Work Dan Roth	15 min.

Textual Entailment for 0-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.

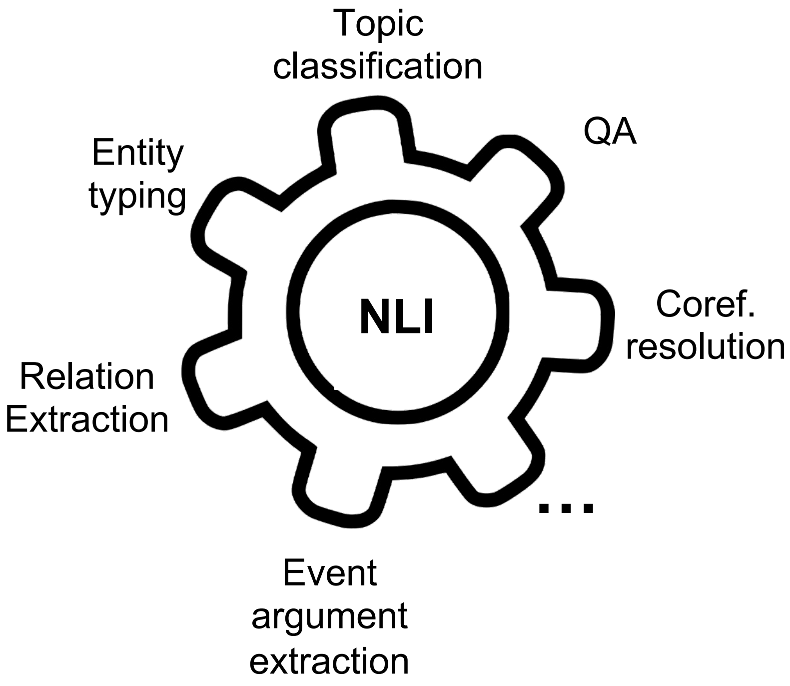
Hypothesis

This text expresses anger.

This text is about accident.

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

Indirect Supervision from Textual Entailment (NLI)



Benefits

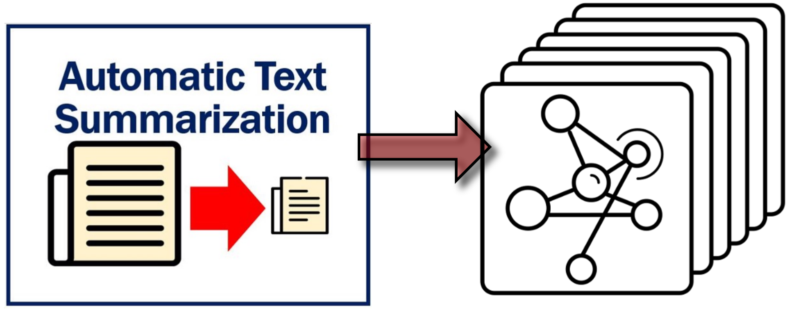
- ❑ Scarce-annotation NLP
- ❑ Cross-task transferability
- ❑ Maximize the potential of small PLMs

Challenges & Solutions

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

Indirect Supervision from Text Generation

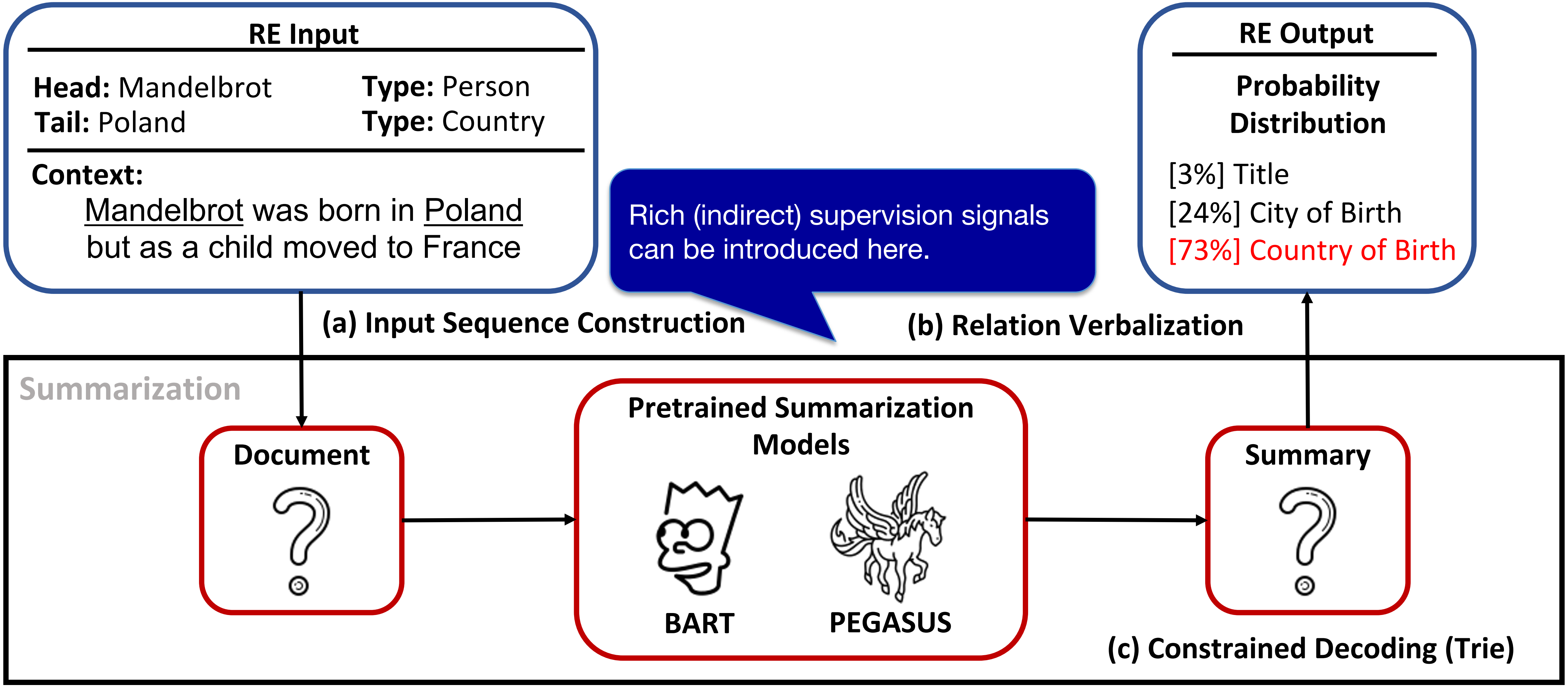
1. Constrained Generation as Indirect Supervision



2. QA as Indirect Supervision

3. IR as Indirect Supervision

Constrained Decoding as Indirect Supervision

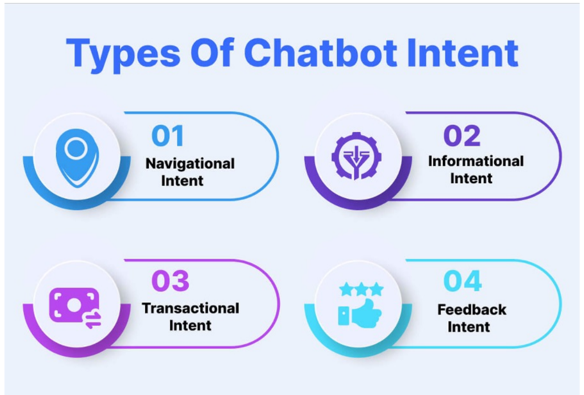


Allowing supervision signals to be transferred from rich summarization resources (CNN/Daily Mail, XSUM) or pretrained models (BART-CNN, Pegasus).

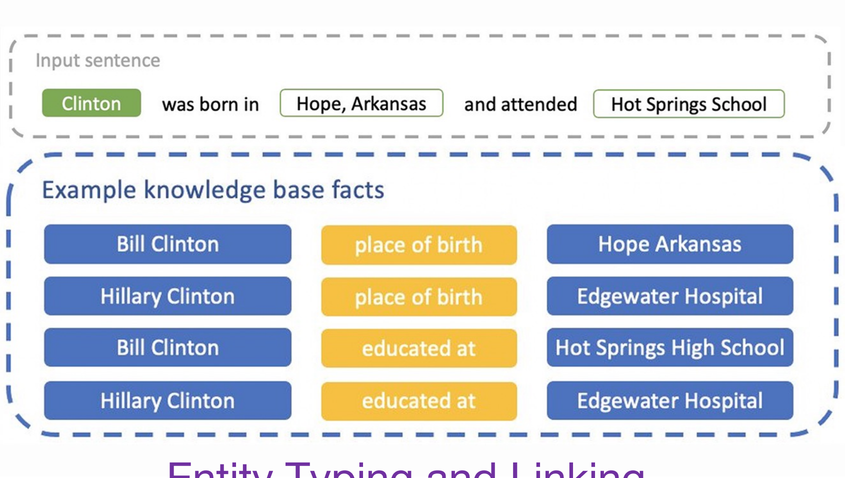
Lu et al. Summarization as Indirect Supervision for Relation Extraction. **EMNLP** 2022

Dense Retrieval as Indirect Supervision for Large-space Decision Making

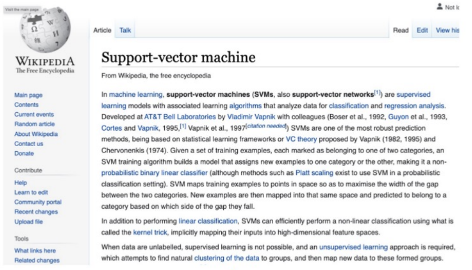
- NLU tasks may have hundreds to millions of decisions



Intent Detection

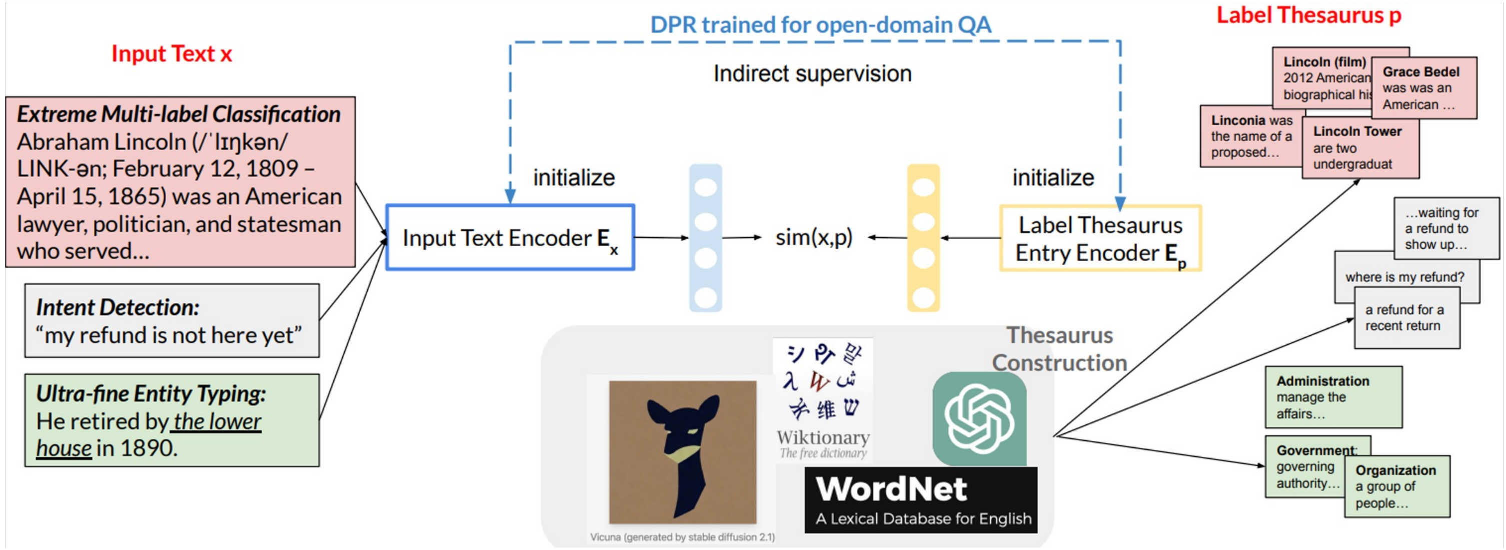


Entity Typing and Linking



Tags:
machine learning
AI
supervised learning

Extreme multi-label classification (XMLC)



Learning to retrieve from a decision thesaurus as a general solution

Natural Text as Supervision

- Natural Texts are structured to contain rich information

  How to generalize beyond the simple-minded pre-training done today?

  Pre-trained language models (LMs) are a great proxy to use NT “incidentally”

  However, they are flawed in a few major ways

  - Cannot accurately capture local relational information (relation type / numbers)
  - Cannot efficiently connect global information (e.g., more than one documents)
  - Large LMs lack controllability without direct supervision (which can be hard to integrate)

  Due to reporting biases, these flaws limit LM’s reasoning capabilities.

- In this section of our tutorial, we discuss

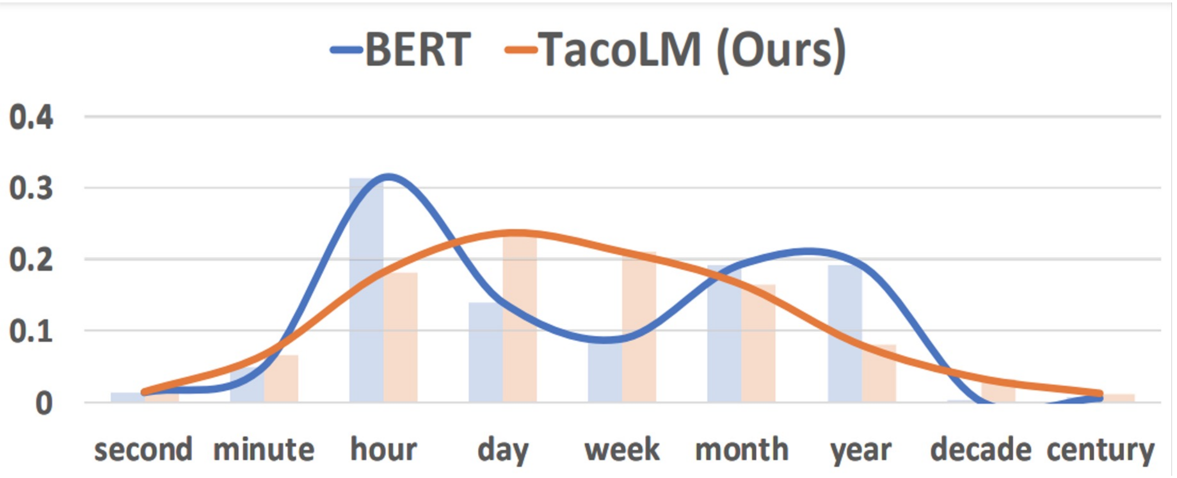
  How local texts can be more efficiently parsed and injected into models

  How to utilize global information from natural texts

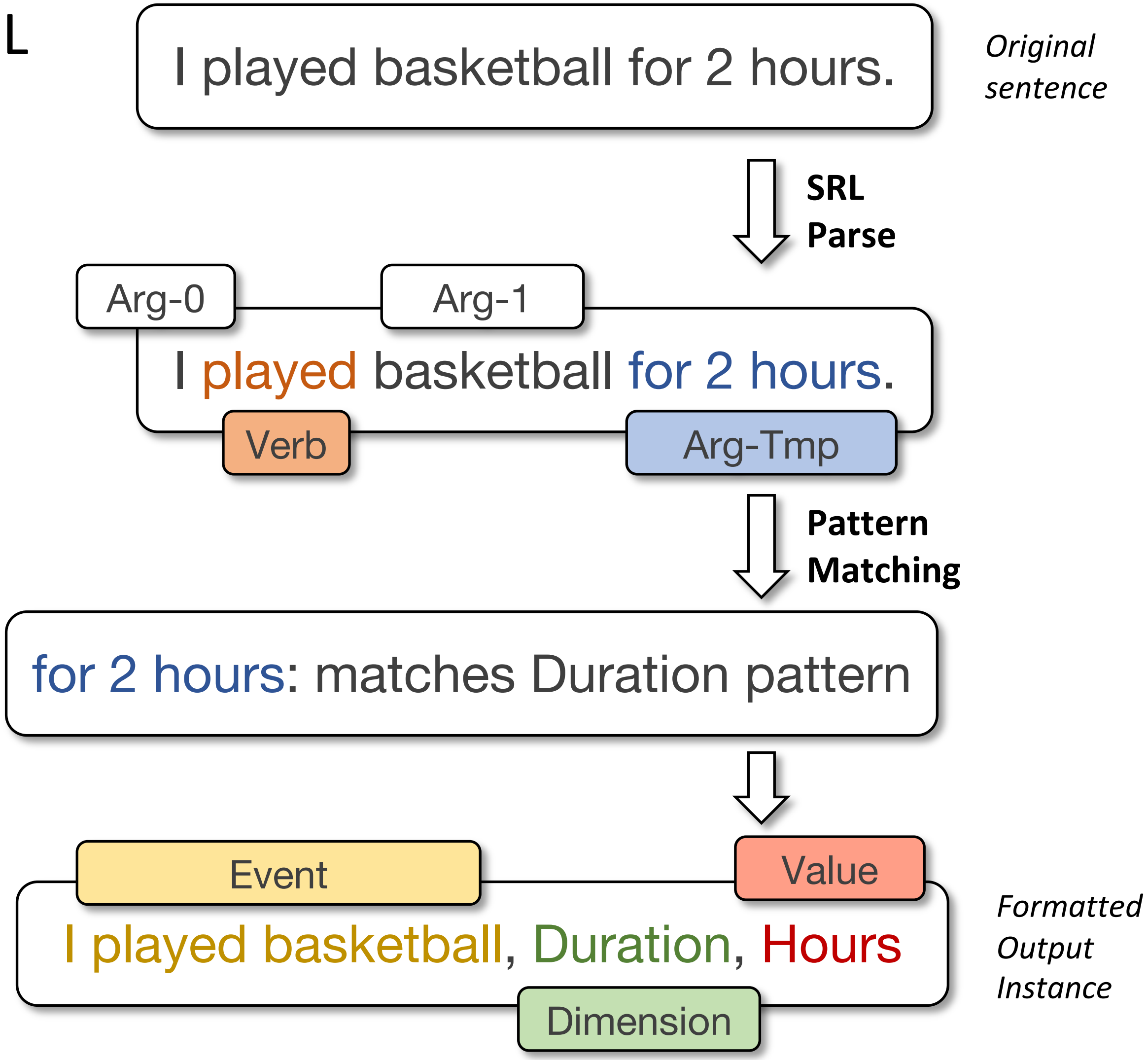
  How LMs can be used to viewed as a generator of incidental signals from NT

Capture Local Information

- Use high-precision patterns based on SRL



Averaged duration prediction on a set of events with gold durations of “days”



Capture Global Information

- **Cross-sentence extraction**
  - Based on explicit temporal expressions
  - Independent of event locations
  - Produces relative distance between start times

text

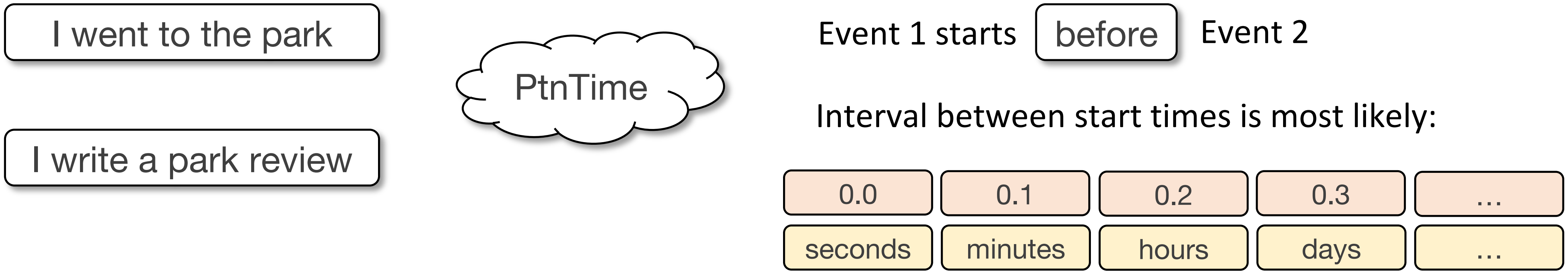
I went to the park on January 1st. I was very hungry after some hiking. Luckily, I purchased a lot of food before I went to the park. I enjoyed the trip and wrote an online review about the trip on the 10th.

within-sentence

[I purchased food, I went to the park.]: **before**

cross-sentence

[I went to the park, I wrote a review]: **before**, weeks

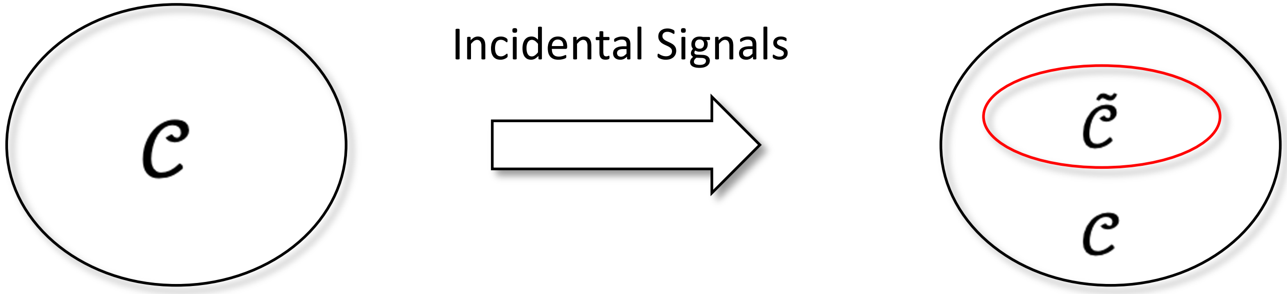


Zhou et al., *Temporal Reasoning on Implicit Events from Distant Supervision*, NAACL 2021

The Impact of Incidental Signals on the Concept Class

Why do incidental signals help learning?

- $c: X \rightarrow Y, where\ c \in \mathcal{C}$
- Learning theory shows that the size of the concept class determines the "easiness" of the learning problem
  - E.g. the generalization bound $R(c) \leq \hat{R}(c) + \sqrt{\frac{\ln|\mathcal{C}|+\ln\frac{2}{\delta}}{2m}}$
- We will show that the use of incidental signals reduces the size of the concept class, and then will use the relative size of the reduction as a measure for the informativeness of the incidental signals



Original Concept Class

Reduced Concept Class

$$S(\mathcal{C}, \tilde{\mathcal{C}}) = \sqrt{1 - \frac{\ln|\tilde{\mathcal{C}}|}{\ln|\mathcal{C}|}}$$

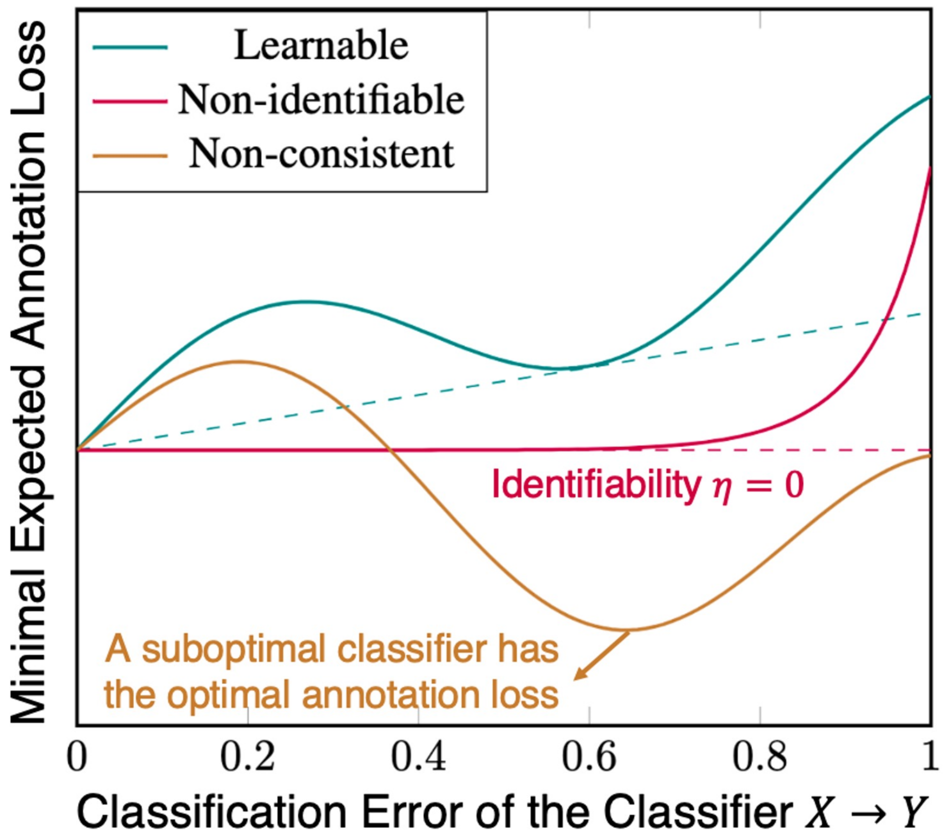
Smaller $\tilde{\mathcal{C}}$ leads to higher Informativeness S

Reduce the concept class from $\mathcal{C}$ to $\tilde{\mathcal{C}}$

Learnability Condition: Overview

To illustrate the learnability condition, we plot the relationship between the classification error of a hypothesis h and the minimum annotation loss (risk) it can have (over choices of transition hypotheses).

Under what conditions are incidental signal sufficient to support learning?



Learning Vision-and-Language Alignments and Representation

Several people **walking** on a **sidewalk** in the **rain** with **umbrellas**.

Main training objective is to predict missing words.

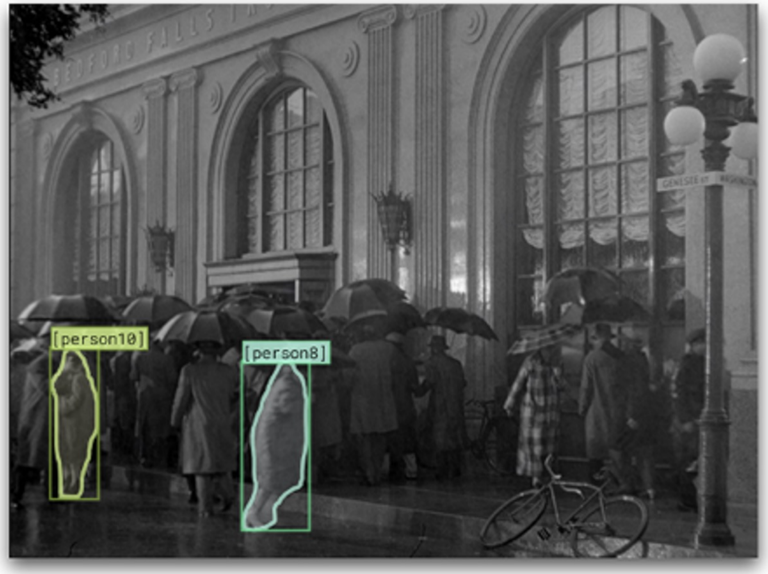
VisualBERT

The model projects words and image regions into the same vector space and uses multiple Transformer layers to build joint representations.

Several people [MASK] on a [MASK] in the [MASK] with [MASK].

Input consists of an image and a caption with some masked words. Such data is easy to obtain from the internet.

Unsupervised pre-training on vision and language



Is it raining outside?

a) Yes, it is snowing.

b) Yes, [person8] and [person10] are outsid

c) No, it looks to be fall.

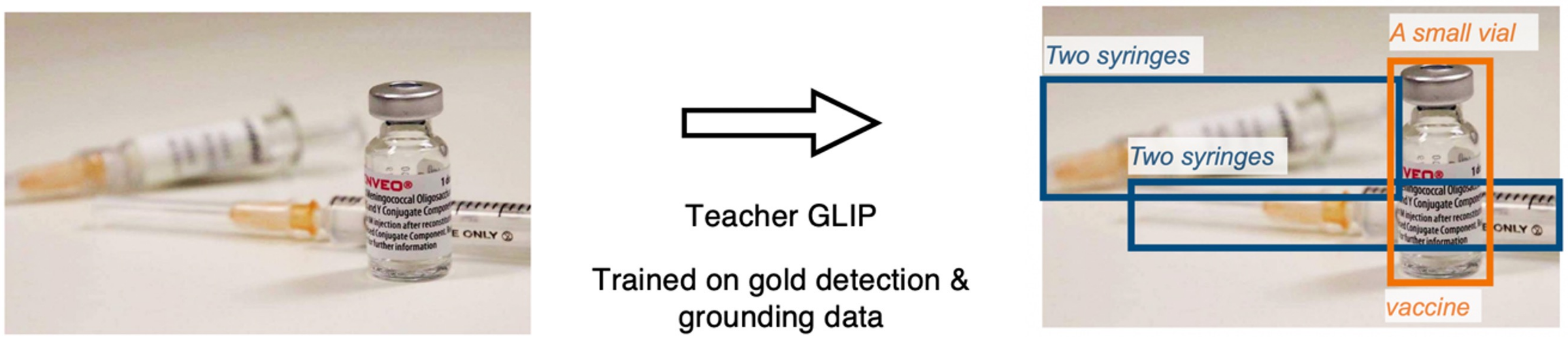
d) Yes, it is raining heavily.

An example from the VCR dataset

Transfer to answering commonsense questio

Generating Pseudo Grounding Data from Captions

- Learn to align objects in image and phrases in text
  - Train a teacher model with gold grounding data; produces boxes given image-caption data
  - Distant supervision assumption: objects in the images are likely to be mentioned in captions

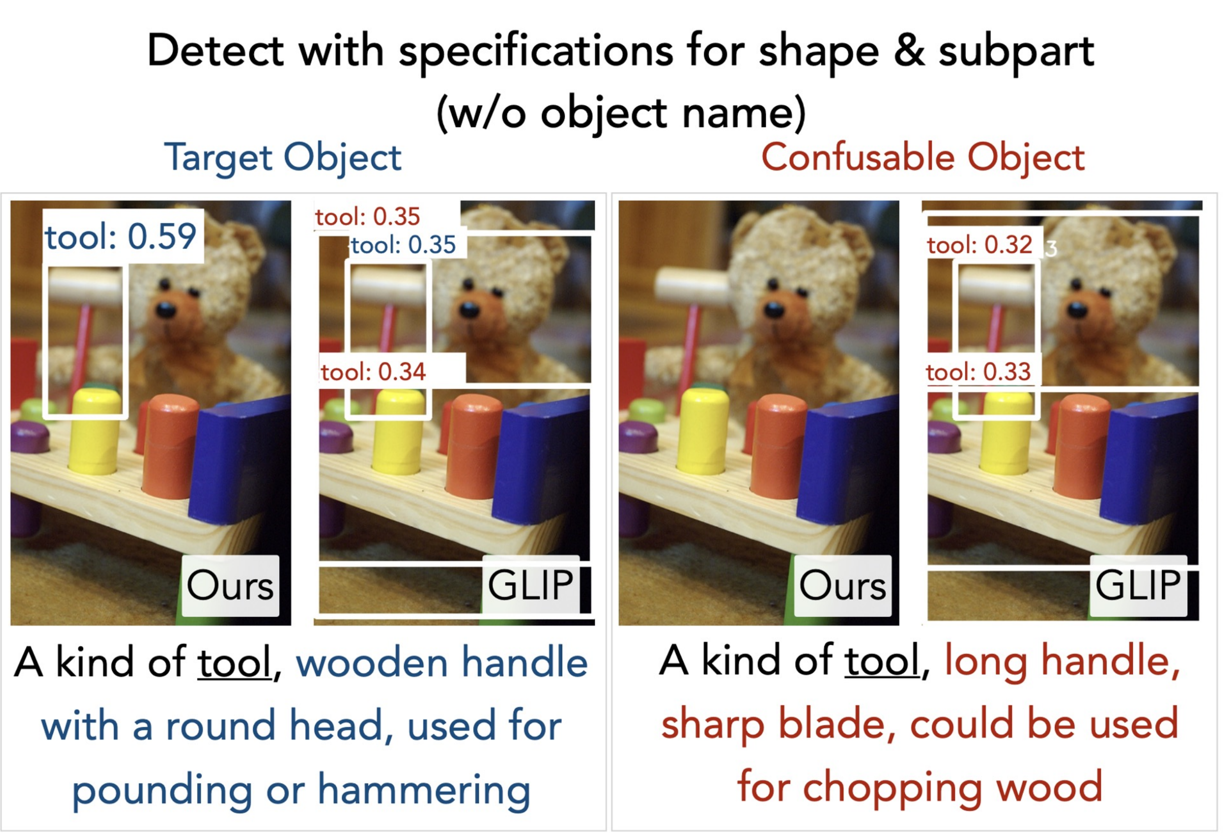


Two syringes and a small vial of vaccine.

Two syringes and a small vial of vaccine.

Learning Visual Concepts from Descriptions

- Using LLM as commonsense engine to specify visual concepts
- Enforcing VL models to align objects with rich description



Tutorial Outline

- Introduction Dan Roth	20 min.
- Indirect Supervision from text classification Wenpeng Yin	30 + 5 min.
- Indirect Supervision from text generation Muhao Chen	30 + 5 min.
- Break	30 min.
- Incidental Supervision from Natural Text Ben Zhou	25 + 5 min.
- Theoretical Analysis of Incidental Supervision Qiang Ning	25 + 5 min.
- Indirect Supervision from Multi-modalities Kai-Wei Chang	25 + 5 min.
- Conclusion and Future Work Dan Roth	15 min.