Incidental Supervision from Natural Text
Indirectly Supervised Natural Language Processing (Part III)

Ben Zhou
Department of Computer and Information Science
University of Pennsylvania
Natural Texts as Supervision

Natural Texts are structured to contain rich information

- Pre-trained language models (LMs) are a great proxy to use NT “incidentally”
- However, they are flawed in a few major ways
  - 1. cannot accurately capture local relational information (relation type / numbers)
  - 2. cannot efficiently connect global information (e.g., more than one documents)
  - 3. large LMs lack controllability without direct supervision (which can be hard to integrate)
- Because of the reporting biases, these three 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
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

Two examples
- Temporal Common Sense
- Speaker Identification
Temporal Common Sense

Improve numerical representation and relation types

Dr. Porter is **taking a vacation** and **will not** be able to see you soon.

Dr. Porter is **taking a walk** and **will** be able to see you soon.
Challenging Reporting Biases:

- people rarely mention the common sense to be efficient “It took me 2 seconds to move my chair”
- We need to specifically find such information, and use them more efficiently
Use high-precision patterns based on SRL
- Duration
- Frequency
- Typical Time
- Duration Upperbound
- Hierarchy

Labels
- Units (seconds, … centuries)
- Temporal keywords (Monday, January, …)

Output
- 4.3M instances of (event, dimension, value) tuple

SRL Parse

Pattern Matching

for 2 hours: matches Duration pattern

I played basketball, Duration, Hours

Original sentence

I played basketball for 2 hours.

Verb

Arg-0

Arg-1

Arg-Tmp

Event

Value

Dimension

Zhou et al., Temporal Common Sense Acquisition with Minimal Supervision, ACL 2020
Joint Model with Masked LM

- 1. Recover Fine-grained Relations and Accurate Numerical Values
- 2: Soft cross entropy for recovering Val
  For a gold duration label “days”, predicting “hours” is more acceptable than “seconds”
- 3: Label weight adjustment
  Instances with “seconds” have higher loss than those with “years”

Trains a BERT-based model called TacoLM

Zhou et al., *Temporal Common Sense Acquisition with Minimal Supervision*, ACL 2020
A collection of events with duration of “seconds,” “weeks” or “centuries” (three extremes)

- BERT (left), TacoLM (right) representation on these events with 2-D visualization
- TacoLM separates the events much better (⇒ more aware of time)
Evaluation: Intrinsic (Quantitatively)

- Metric: Distance to gold label
  - Dist (seconds, hours)=2, Dist (minutes, hours)=1
  - **Lower the better**
- Annotated Temporal Commonsense Benchmark

```
Duration | Frequency | Typical Time (avg)
---------|-----------|-------------------
BERT     | TacoLM    |                   
1.33     | 1.68      | 1.98              
0.75     | 1.17      | 1.74              
```

19% average improvement

---

Zhou et al., *Temporal Common Sense Acquisition with Minimal Supervision*, ACL 2020
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

Two examples
- Temporal Common Sense
- Speaker Identification
Speaker Identification

- Speaker Identification (SI): who said which utterances in novels/stories.

- Identify who said each utterances in text

  Alice made a mistake and she wanted to apologize to Jane. “I won’t do it again.” “It’s fine, don’t worry about it.”

- Traditionally viewed as an information extraction task
  
  Semantic role labeling
  Pronoun resolution
  Gender extraction

No coreference
No gender
No alternation

Existing Supervision only annotates instances with direct evidences, so we need more diverse cases from incidental supervision

Zhou et al. "Cross-Lingual Speaker Identification Using Distant Supervision" 2022
Find Incidental Supervision

- **IE-based speaker identification**

  Alice made a mistake and she wanted to apologize to Jane. “I won’t do it again,” she said. “It’s fine, I forgive you” Jane said.

- **Direct Speaker Identification**
Find Incidental Supervision

- IE-based speaker identification
  
  Alice made a mistake and she wanted to apologize to Jane. “I won’t do it again,” she said. “It’s fine, I forgive you” Jane said.

- Direct Speaker Identification

- Conversation Alternation Patterns

Zhou et al. "Cross-Lingual Speaker Identification Using Distant Supervision" 2022
He et al. "Identification of Speakers in Novels." 2013
Find Incidental Supervision

- IE-based speaker identification

Alice made a mistake and she wanted to apologize to Jane. “I won’t do it again,” she said. “It’s fine, I forgive you” Jane said.

- Direct Speaker Identification
- Conversation Alternation Patterns
- Local Coreference Resolution
Reasoning with Incidental Supervision

- Our IE pipeline relies on “explicit” clues to find speakers
  It will not encourage contextual reasoning

  Alice made a mistake and she wanted to apologize to Jane. “I won’t do it again,” she said. “It’s fine, I forgive you” Jane said.

- Randomly remove explicit direct speaker mentions
  The model must use the context to figure out the speakers
Experiments on Speaker Identification

- Pride & Prejudice Dataset

- DISSI outperforms previous supervised method (+5%) without supervision

Zhou et al. "Cross-Lingual Speaker Identification Using Distant Supervision" 2022
He et al. "Identification of Speakers in Novels." 2013
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

Two examples
- Temporal Relation
- Question Decomposition
PatternTime: Distant Supervision Collection

- We want to learn to compare **start times**
  - From unannotated free texts
- **Within-sentence extraction**
  - Not enough:
    - LM can easily learn such relations
    - Does not address implicit events
    - Does not tell how far the two start times are

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.

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

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

Zhou et al., Temporal Reasoning on Implicit Events from Distant Supervision, NAACL 2021
We want to learn to compare start times from unannotated free texts.

Cross-sentence extraction

Based on explicit temporal expressions

Independent of event locations

Produces relative distance between start times.

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

PatternTime

- A sequence-to-sequence model
  - Train on 1.5M distant supervision instances
- Input: two event phrases
- Output:
  - A binary label indicating which event starts earlier
  - Probabilities over duration units indicating the interval between two start times

I went to the park

I write a park review

Event 1 starts before Event 2

Interval between start times is most likely:

0.0 seconds 0.1 minutes 0.2 hours 0.3 days ...

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

- On TRACIE dataset (from the same paper)
  Evaluates event temporal relations (both start time and end time comparison)
  All models/baselines are trained with TRAICE training set

![Bar chart showing performance improvements for T5-Large, PatternTime, and T5-3B models on TRACIE dataset.

Incidental supervision brings significant performance improvements.
Comparison of within-sentence / cross-sentence

TRACIE start time accuracy

- Cross-sentence (global information) contributes the most since it introduces new information to LMs
Experiments: TRACIE

- When training data has different gold label distribution
- Same test set (lower the better)

Incidental supervision helps to produce stable model that is less affected by supervision biases.

Zhou et al., Temporal Reasoning on Implicit Events from Distant Supervision, NAACL 2021
Ning et al., A Multi-Axis Annotation Scheme for Event Temporal Relations, ACL 2018
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

Two examples
- Temporal Relation
- Question Decomposition
Decomposition Reasoning with Incidental Sup.

- Reasoning can be viewed as finding equivalencies that suit best for a reasoner.

  Can Cyril Ramaphosa become Secretary General of NATO?

  A geopolitical expert reasoner
  No.

  Cyril Ramaphosa is the president of South Africa. NATO only contain Atlantic countries.

  Oh OK, so no.

  An educated adult reasoner
  Idk.

  NATO is an organization of countries. Its Secretary General is selected among leaders of membership countries. Cyril Ramaphosa is the leader of South Africa, which is not part of NATO.

  No.

  A 6-year-old reasoner
  I don’t want to go to school
  Still not sure

  No

Zhou et al., Learning to Decompose: Hypothetical Question Decomposition Based on Comparable Texts, EMNLP 2022
How Should We Decompose?

- Decomposition is about finding equivalent reasoning processes with respect to a goal.
- Why existing models struggle to find these equivalencies?
  Reporting bias: authors do not repeat a process with another equivalent one
  Language models cannot easily pick up such equivalencies

How do we mitigate such a gap?
With **Incidental Supervision**
Learn to decompose from **comparable texts**
Parallel news articles that **describe the same things from different angles**

- **Document A**
  - The Albany in NY is more crowded than that in GA.
  - While they are prevalent today...
  - ...latest environment protection...

- **Document B**
  - The Albany in NY has more people and less space.
  - There is a large number of these...
  - The administration... reducing methane gas...

**Is cow methane safer for the environment than cars?**

We need to compare the quantity of methane gas, lower the safer.

---

Zhou et al., *Learning to Decompose: Hypothetical Question Decomposition Based on Comparable Texts*, EMNLP 2022
Evaluating **DECOMPT5**

- **Overnight**
  - Hit@K accuracy

\[
\begin{align*}
&\text{Hit@1} \\
&\text{Hit@5} \\
&\text{Hit@10}
\end{align*}
\]

- **T5**
- **DecompT5 (ours)**

Hit@K accuracy for Overnight evaluation.
Evaluating DeCOMPT5

- TORQUE

Exact match accuracy

T5-paraphrasing: a baseline trained with distant paraphrasing signals
A QA pipeline that uses DecompT5 for question decomposition
Evaluating DECOMPENTAIL

- On HotpotQA

![HotpotQA (binary) acc. Comparison](image)

- T5-Large
- GPT-3
- GPT-3 CoT
- DecompEntail
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
Temporal Reasoning is inherently challenging for LMs
  Reporting biases + numerical issues
Recall PatternTime: an incidentally-supervised T5 for temporal reasoning

Incidentally-supervised SLM is much better than few-shot LLM on almost all temporal datasets
If LLM is not very good at such tasks, can we still utilize its semantic understanding?

We introduce how we can use LLM to generate

- Incidental training instances
- Incidental explanations for better inference

If we can select the good ones!
Temporal reasoning as an example

- Temporal differential analysis (at ACL 2023)

I only took lunch today while my parents had both lunch and dinner.

Original Context, Event 1 (lunch) and Event 2 (dinner)

My parents are traveling in China, and I am in the states.

Extra Context (additional sentence)

Evaluates: Does the extra context makes Event 1 more BEFORE Event 2, or more AFTER?

Existing temporal datasets only annotate “hard” labels, which will mark “lunch” to be before “dinner”. However, the current context is inconclusive.

Since China’s time zone is ahead of the States, this increases the likelihood of “dinner” before “lunch”
Generating Incidental Supervision

- Today dataset: Annotates 1,241 training examples with event pairs, contextual change as additional sentences, and explanations
  - Expensive to annotate, Not enough to supervise certain models
  - Can we use the semantic power of LLMs to generate more?

I met Ben at the coffee shop in the morning, who just finished a meeting.

I woke up in the morning

Ben’s meeting started

Can you add a sentence to make this temporal relation more “before”?

I went to the park first thing in the morning. ✓

Ben had a long meeting this morning. ✗

Multiple ways can be used to filter generated instances

Feng et al., Generic temporal reasoning with differential analysis and explanation, ACL 2023
Several SLMs can be trained to mitigate different sources of mistakes from LLM-generated instances

Temporal relation prediction disagreement between SLMs and LLMs with generated additional sentence and explanations
Seemingly convincing explanations but incorrect additional sentence + label
Seemingly correct additional sentence + label, but incorrect explanations

Human-designed heuristics are also helpful
E.g., any additional sentence repeating the original context is bad
Training with LLM-generated instances

- Annotated Supervision
  - 1,214 Today examples
  - 1,500 Matres examples
  - 860 Tracie examples

- Incidental Supervision
  - 5000 GPT-3.5 generated instances
  - 1,475 after filtering

Feng et al., *Generic temporal reasoning with differential analysis and explanation*, ACL 2023
Zhou et al., *Temporal Reasoning on Implicit Events from Distant Supervision*, NAACL 2021
Ning et al., *A Multi-Axis Annotation Scheme for Event Temporal Relations*, ACL 2018
Experiments with Incidental Training Instances

- Base model: T5-large

![Diagram showing comparison between Annotated and Annotated+Incidental models]

- TODAY
- TODAY (gold explanation)
LLMs can provide explanations or “reasons” that are semantically relevant to the task. SLMs can benefit from these explanations to act better on filtering and decision.
Future Directions

Post-hoc verifications for LLMs with incidental signals from natural text

- Incidentally-supervised Small Language Models
- Symbolic Data/Knowledge Base
- Controlled Inference
- Finetune / RLHF

LLM-guided incidental supervision from natural text

- Natural Texts
- Large Language Models
- Semantic Abstraction
- Linguistic Pattern Extension
In this part of the tutorial, we show that

Pre-trained language models are inherently limited by the way they acquire information from natural text. We can get more information by

- Establishing clear local connections
- Build long-distant and global relations

Moreover, large language models provide strong semantic correlations, but could fail on complicated tasks (e.g., temporal reasoning). We can view such semantic correlations as signals from natural texts, and augment supervised smaller models with

- Incidental Training instances
- Incidental Explanations

Thank you!