Indirect Supervision from Generative and Retrieval Tasks
Indirectly Supervised Natural Language Processing (Part II)

Muhao Chen
Department of Computer Science
University of Southern California

July 2023
ACL Tutorials
Indirectly Supervised Natural Language Processing
How do we support more *expensive* NLU tasks with more *resource-rich* NLG/IR tasks?
The Root of All Problems: Expensive Supervision

Obtaining direct supervision can be **difficult and expensive**

- Reading long documents, recognizing complex structures
- Costly effort from expert annotators

**Insufficiency**

- **General domain**: A few hundred documents or ten thousand scale sentences with annotation
- **Specialized domains**: Up to several thousand sentences.

**Low-resource Domains with Almost No Annotations**

- ~$7 per label in the general domain [Paulheim, 2018].
- ~$71 per label in proteomics domain [Sullivan+, 2017].
- Even more unaffordable for drugs, diseases, clinical trials …

Result: Poor Generalization
Challenge: Conditioned Decision Making

NER / Event Extraction (Conditioned on surrounding tokens)

Relation Extraction (Conditioned on entity mentions)

Hard to be modeled as NLI
- Diverse preconditions in the same context (different spans, entity pairs in the same input)
- Ambiguous entailment (the same input needs to entail many hypotheses)
Challenge: Very Large Decision Spaces

Entity Linking, Fine-grained Typing

Tasks with **very large decision spaces** that to be supervised as NLI or classification.

Thousands to millions of labels, more like a dictionary.
Non-discriminative or structured decisions that are beyond the ability of NLI

Passage Context

The Matrix is a 1999 science fiction action film written and directed by The Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix": created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world."

Question

Who stars in The Matrix?
In This Talk

1. Constrained Generation as Indirect Supervision
2. QA as Indirect Supervision
3. IR as Indirect Supervision
Insufficient Structural Annotations

Information extraction suffers from insuffcient supervision

Direct annotation is difficult and expensive
Can we transfer signals from a more resource-rich task?

An Exemplary Form of Indirect Supervision

Summarization as Indirect Supervision
Take Relation Extraction As An Example

**Context (Sentence)**

Mandelbrot was born in Poland but as a child moved to France

**Head Entity**

Benoit Mandelbrot

**Tail Entity**

Poland

**Head Mention**

Mandelbrot

**Tail Mention**

Poland

**Relation?**

Title, City of Birth, Country of Birth

Formulated As Multi-class Classification

- Heavily relying on enough relation annotations

Almost cannot generalize to rarely seen or unseen relations.

- LUKE (Full-shot)
- LUKE (1% training data)
Indirect Supervision from Abstractive Summarization

**Summarization:** Generating concise expressions of *synoptical information* from the longer context

**Document**

Authorities said the incident took place on Sao Joao beach in Caparica, south-west of Lisbon. The National Maritime Authority said a middle-aged man and a young girl died after they were unable to avoid the plane. [6 sentences with 139 words are abbreviated from here.] Other reports said the victims had been sunbathing when the plane made its emergency landing. [Another 4 sentences with 67 words are abbreviated from here.]

**Summary**

A man and a child have been killed after a light aircraft made an emergency landing on a beach in Portugal.

**A more resource-rich task**

- **Million-scale parallel summary corpora** (vs. a few hundred docs or <100k sentences for RE)
- More **easy-to-consume sources** (news summaries, paper abstracts, etc.)

Relation is just **one kind of synoptical information**. Can we reformulate RE as summarization?
Reformulating RE as Summarization

**RE Input**

- **Head:** Mandelbrot
- **Type:** Person
- **Tail:** Poland
- **Type:** Country

**Context:**
Mandelbrot was born in Poland but as a child moved to France

**RE Output**

- **Probability Distribution**
  - [3%] Title
  - [24%] City of Birth
  - [73%] Country of Birth

**Summarization**

- **(a) Input Sequence Construction**
- **(b) Relation Verbalization**
- **(c) Constrained Decoding (Trie)**

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
Rewriting Inputs and Outputs

Input Sequence Construction

- **Adding entity mentions and types**: hint the summarization model which entity pair is targeted for summarization.

Relation Verbalization

- Simple template-based verbalization (using surface names of relations)

<table>
<thead>
<tr>
<th>Relations</th>
<th>Templates</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_1$ title</td>
<td>$T_1$ {subj} is a {obj}</td>
</tr>
<tr>
<td>$r_2$ city of birth</td>
<td>$T_2$ {subj} was born in the city {obj}</td>
</tr>
<tr>
<td>$r_3$ country of birth</td>
<td>$T_3$ {subj} was born in the country {obj}</td>
</tr>
<tr>
<td>$r_4$ founded by</td>
<td>$T_4$ {subj} was founded by {obj}</td>
</tr>
<tr>
<td>$r_5$ NA</td>
<td>$T_5$ {subj} has no known relations to {obj}</td>
</tr>
</tbody>
</table>

Both become natural language text that fits a summarization model.

Lu et al. Summarization as Indirect Supervision for Relation Extraction. EMNLP 2022
Training Process: Transfer Finetuning A Summarization Model

Summarization (Training)

Document
The subject entity is Mandelbrot. The object entity is France. The type of Mandelbrot is person. The type of France is country. Mandelbrot was born in Poland but as a child moved to France.

Pretrained Summarization Models
BART
PEGASUS

Summary
Mandelbrot was born in the country Poland

Decoding

Inference: How to make sure the model only generates relation information? (i.e. mappable to the decision space of RE)

RE Output
Probability Distribution
[3%] Title
[24%] City of Birth
[73%] Country of Birth

Lu et al. Summarization as Indirect Supervision for Relation Extraction. EMNLP 2022
Inference with Trie-based Constrained Decoding

How to let the model summarize only relation-descriptive information?

**Step 1.** Build a Trie for relations

**Step 2.** Beam Search on the Trie

**Step 3.** Calculate accumulate scores

Next Token Prediction in Beam Search becomes Next Forky Node Prediction.

Lu et al. Summarization as Indirect Supervision for Relation Extraction. EMNLP 2022
Summarization provides strong indirect supervision for low-resource relation extraction. Also leads to precise full-shot relation extraction.

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

Input: <s> have muscle pain and fatigue </s>
Output: 2 3 7 2 5 6
Earlier Monday, a 19-year-old Palestinian riding a bicycle detonated a 30-kilo (66-pound) bomb near a military jeep in the Gaza Strip, injuring three soldiers.

Hsu et al. DEGREE: A Data-Efficient Generation-Based Event Extraction Model. NAACL 2022
Agenda

1. Constrained Generation as Indirect Supervision
2. QA as Indirect Supervision
3. IR as Indirect Supervision
Two Forms of QA as Generalizable Indirect Supervision

**Extractive [SQuAD]**

**Question:** At what speed did the turbine operate?
**Context:** (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...
**Gold answer:** 16,000 rpm

**Extractive QA**
Supporting decisions inclusive to the input text
- Span detection (NER, Coref, etc.)
- Parsing (SRL, AMR, etc.)

**Abstractive [NarrativeQA]**

**Question:** What does a drink from narcissus's spring cause the drinker to do?
**Context:** Mercury has awakened Echo, who weeps for Narcissus, and states that a drink from Narcissus's spring causes the drinkers to "Grow dotingly enamored of themselves." ...
**Gold answer:** fall in love with themselves

**Abstractive/Generative QA**
Supporting any free-form decisions
- Relation extraction
- Dialogue
- Intent prediction
- etc.

Span or structural decisions.

Free-form decisions
**Benefit 1: Handling nested entity mentions (not feasible for sequence tagging)**

Last night, at *the Chinese embassy in France*, there was a holiday atmosphere.

**Benefit 2: Questions serve as label definitions**  
(Further improving generalization)

<table>
<thead>
<tr>
<th>Entity</th>
<th>Natural Language Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>Find locations in the text, including non-geographical locations, mountain ranges and bodies of water.</td>
</tr>
<tr>
<td>Facility</td>
<td>Find facilities in the text, including buildings, airports, highways and bridges.</td>
</tr>
<tr>
<td>Organization</td>
<td>Find organizations in the text, including companies, agencies and institutions.</td>
</tr>
</tbody>
</table>

Better performance than seq2seq generation

![System F1](chart.png)

Li et al. A Unified MRC Framework for Named Entity Recognition. ACL 2020
Using the **sentence that each mention is in** as the “question”, **all other spans** belonging to the same cluster as “answers”
Other Tasks as Extractive QA

---

**QA-SRL: QA as Semantic Role Labeling**

<table>
<thead>
<tr>
<th>Relation</th>
<th>Question</th>
<th>Sentence &amp; Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>educated at</td>
<td>What is Albert Einstein’s alma mater?</td>
<td>Albert Einstein was awarded a PhD by the University of Zürich, with his dissertation titled...</td>
</tr>
<tr>
<td>occupation</td>
<td>What did Steve Jobs do for a living?</td>
<td>Steve Jobs was an American business, inventor, and industrial designer.</td>
</tr>
<tr>
<td>spouse</td>
<td>Who is Angela Merkel married to?</td>
<td>Angela Merkel’s second and current husband is quantum chemist and professor Joachim Sauer, who has largely...</td>
</tr>
</tbody>
</table>

---

**Advantages of Extractive QA for Information Extraction Tasks (over Seq2Seq Gen)**

- Handling nested spans
- Questions can serve as task-oriented prompts and semantic representation of the label space

---

Fitzgerald et al. Large-Scale QA-SRL Parsing. ACL 2018  
Benefits of An Abstractive QA Reformulation

- Supporting free-form, non-discriminative decision making
- Supporting multiple answers

Task-oriented Parsing (e.g., predicting user intent)

Event argument generation

Q: All possible intents from a user are [...], and slots could be [...]. A user said, “Look up directions to the nearest parking near S Beritania Street.” What did the user intend to do?
A: The user intended to get directions, where destination is nearest parking near S Beritania Street. The intent for “nearest parking near S Beritania Street” is to get location, where location’s category is parking and location modifiers are near S Beritania Street; nearest. The intent for “near S Beritania Street” is get location, where location is S Beritania Street and search radius is near.
Generalizability and lack of annotations are more significant challenges here for software version compatibility detection.


Extracting drug-drug interaction

Xu et al. Can NLI Provide Proper Indirect Supervision for Low-resource Biomedical Relation Extraction? ACL 2023

Ma et al. DICE: Data-Efficient Clinical Event Extraction with Generative Models. ACL 2023
Agenda

1. Constrained Generation as Indirect Supervision
2. QA as Indirect Supervision
3. IR as Indirect Supervision
Some NLU tasks may have very large decision spaces

**Types Of Chatbot Intent**
- 01 Navigational intent
- 02 Informational intent
- 03 Transactional intent
- 04 Feedback intent

**Example knowledge base facts**

<table>
<thead>
<tr>
<th>Entity</th>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill Clinton</td>
<td>place of birth</td>
<td>Hope Arkansas</td>
</tr>
<tr>
<td>Hillary Clinton</td>
<td>place of birth</td>
<td>Edgewater Hospital</td>
</tr>
<tr>
<td>Bill Clinton</td>
<td>educated at</td>
<td>Hot Springs High School</td>
</tr>
<tr>
<td>Hillary Clinton</td>
<td>educated at</td>
<td>Edgewater Hospital</td>
</tr>
</tbody>
</table>

**Intent Detection**
- Target: Hundreds of intent types

**Entity Typing and Linking**
- Target: entities in a whole KB

**Extreme multi-label classification (XMLC)**
- Thousands to millions of tags

**Large decision spaces in a hundred- to million-scale**
Supervising a classifier is not ideal

- Too few instances per class
- Meaningless class label representation
- Not generalizable to unseen classes

Learning to lookup a label thesaurus should be more feasible

- A plausible source of indirection supervision: Dense Retrievers
- Meaningful label representation
- Generalizes well to unseen labels
A dual-encoder model for retrieving passages most relevant to a question
- Two encoders $P$ and $Q$ for passages and questions
- Contrastive learning to maximize $P^T Q$ for correct question-passage pair
- Efficient (using MIPS) and generalizable retrieval
Dense Retrieval for Entity Linking

The passage as query

\[ Q = \text{enc}_Q([\text{CLS}]p[\text{SEP}]\psi(\text{topic}(x))) \]

Entity candidates with descriptions

\[ P = \text{enc}_P([\text{CLS}]\phi\text{title}(e) \oplus \phi\text{desc}(e)[\text{SEP}]) \]

Reformulating entity linking into open-domain QA

1. The **retriever** finds top-\(K\) candidate entities mentioned in the passage
2. The **reader** extracts spans of each selected entity

A pre-existing inductive bias that helps retrieve the identities of entities

- 85.8 in-domain micro \(F_1\) and 60.5 out-of-domain \(F_1\)

Zhang et al. EntQA: Entity Linking as Question Answering. ICLR 2022
Indirect supervision for retrieving from a fine-grained pool of intents
Enhancing few-shot generalizability (+5.22~8.50% in accuracy for 5-shot prediction)
Dense retrieval from a decision thesaurus for any large-space decision making tasks

Ways of decision representation

- Lexical knowledge bases (WordNet, Wiktionary),
- LLM generated explanations
- Task training data

Retrieval tasks as a general form of indirect supervision
Dense Retrieval as A General Solution

Indirect Supervision Improves Three Large-space Decision Making Tasks

1. Extreme Multi-label Classification (XMC)

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision @1</th>
<th>Precision @3</th>
<th>Precision @5</th>
<th>Recall @1</th>
<th>Recall @3</th>
<th>Recall @5</th>
</tr>
</thead>
<tbody>
<tr>
<td>MACLR</td>
<td>0.57</td>
<td>12.26</td>
<td>12.50</td>
<td>23.84</td>
<td>28.41</td>
<td>35.47</td>
</tr>
<tr>
<td>DDR</td>
<td>27.78</td>
<td>19.98</td>
<td>15.66</td>
<td>15.03</td>
<td>28.87</td>
<td>35.47</td>
</tr>
</tbody>
</table>

If Clinton maintains his lead in the polls, **he** will be the first Democrat since Franklin D. Roosevelt to be elected to a second full term.

2. Ultra-fine Semantic Typing

- **politician**: a leader engaged in civil administration
- **campaigner**: a politician who is running for public office
- **person**: a human being

3. Few-shot Intent Detection

Xu et al. Dense Retrieval as Indirect Supervision for Large-space Decision Making, 2023
## Pros and Cons of Different IS Sources

<table>
<thead>
<tr>
<th>Sources</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLI</td>
<td>• Generalizable reasoning abilities</td>
<td>• Cannot handle diverse preconditions in the same context</td>
</tr>
<tr>
<td></td>
<td>• Applicable to any (incl. simple) classifiers</td>
<td>• Cannot handle non-discriminative or structured tasks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• High inference cost</td>
</tr>
<tr>
<td>Summarization</td>
<td>• Suitable for tasks that <strong>refine input information</strong></td>
<td>• Less suitable for tasks that need more induction</td>
</tr>
<tr>
<td>Extractive QA</td>
<td>• Can handle <strong>span detection tasks</strong></td>
<td>• Decisions must be inclusive to the inputs</td>
</tr>
<tr>
<td></td>
<td>• Supports <strong>nested spans</strong></td>
<td></td>
</tr>
<tr>
<td>Abstractive QA</td>
<td>• Can handle <strong>free-form decisions</strong></td>
<td>• Less effective in tasks where decisions are inclusive to the inputs (e.g. span detection or sequence tagging)</td>
</tr>
<tr>
<td>Dense Retriever</td>
<td>• Suitable for <strong>large decision spaces</strong></td>
<td>• Not suitable for tasks where decisions are inclusive to the inputs</td>
</tr>
<tr>
<td></td>
<td>• <strong>Efficient</strong></td>
<td></td>
</tr>
</tbody>
</table>
Thank You