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

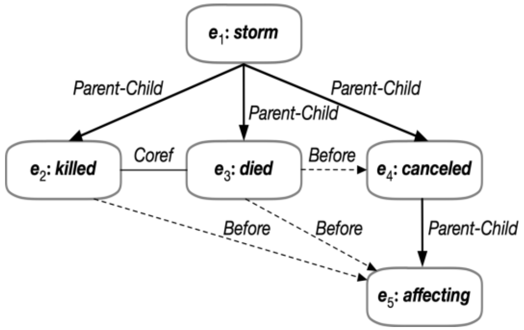
Indirect Supervision from Generative Tasks

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 is **difficult** and **expensive**

On Tuesday, there was a typhoon-strength (e_1:***storm***) in Japan. One man got (e_2:***killed***) and thousands of people were left stranded. Police said an 81-year-old man (e_3:***died***) in central Toyama when the wind blew over a shed, trapping him underneath. Later this afternoon, with the agency warning of possible tornadoes, Japan Airlines (e_4:***canceled***) 230 domestic flights, (e_5:***affecting***) 31,600 passengers.



- Reading long documents, recognizing complex structures

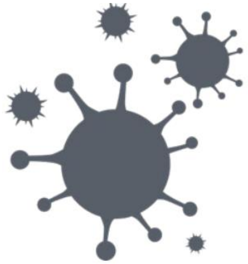
- ~$7 per label in the general domain [Paulheim, 2018].
- ~$71 per label in proteomics domain [Sullivan+, 2017].
- Even more unaffordable for drugs, diseases, clinical trials …

- Costly effort from expert annotators

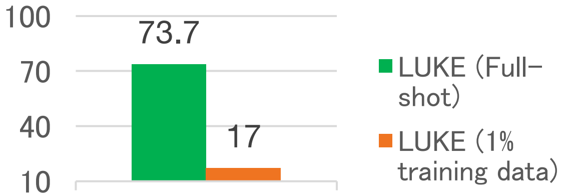
Insufficiency

- **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

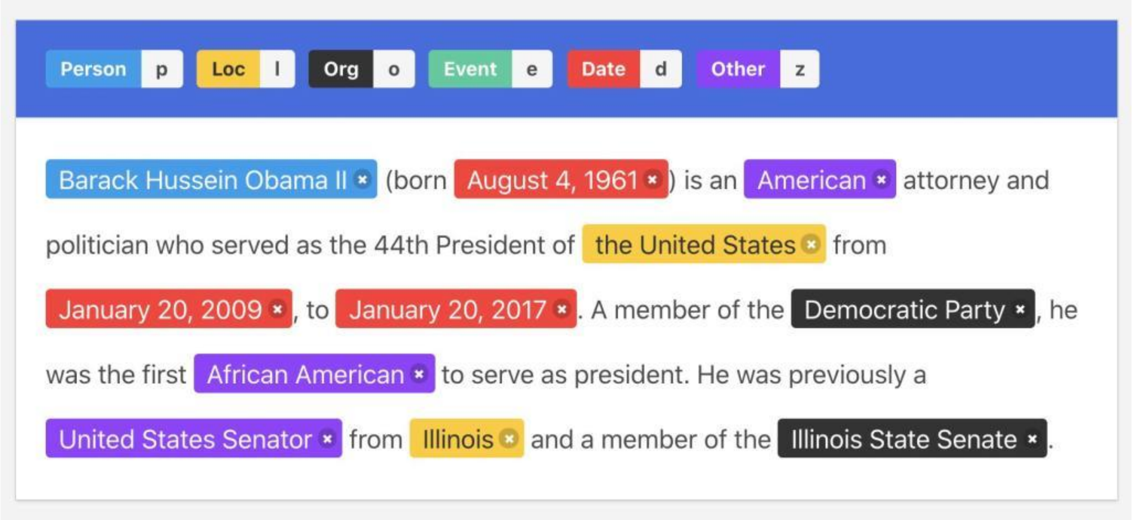


Result: Poor Generalization

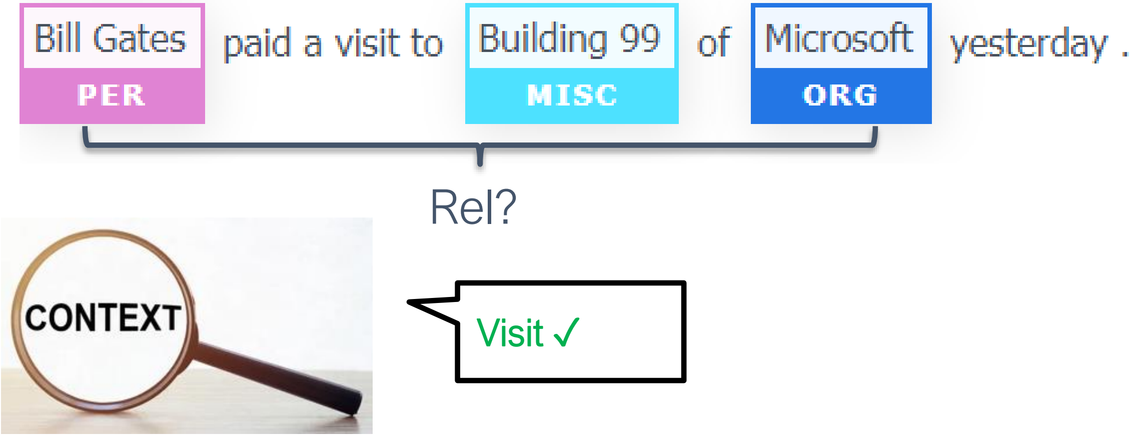


Challenge: Conditioned Decision Making

NER / Event Extraction (Conditioned on surrounding tokens)



Relation Extraction (Conditioned on entity mentions)

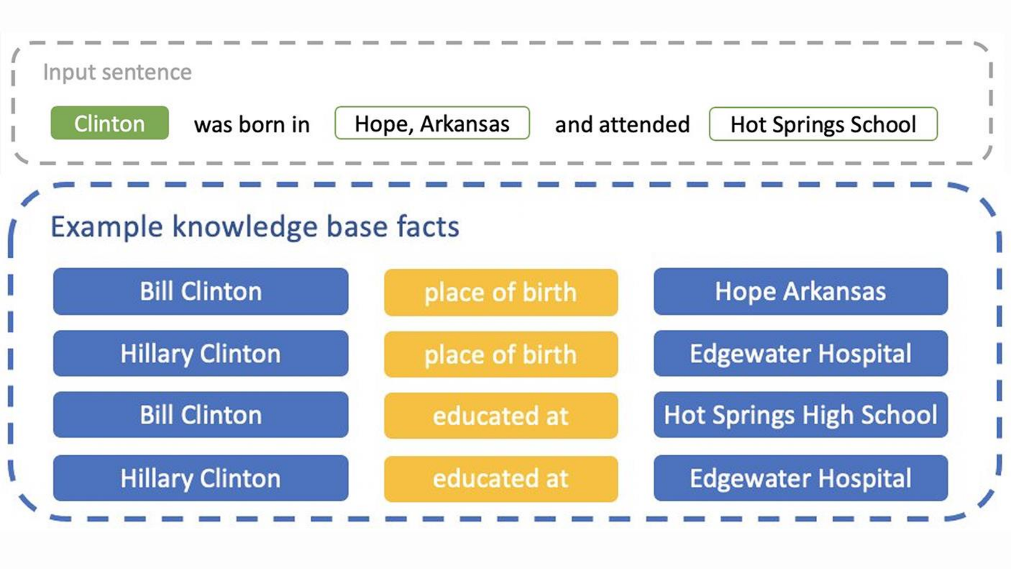


Hard to be modeled as NLI

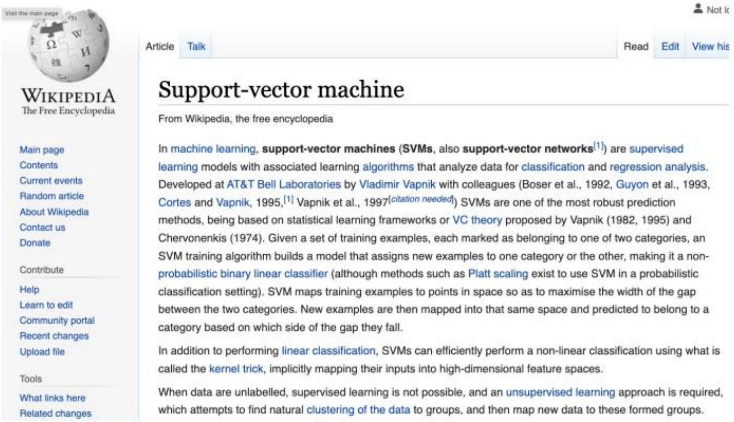
- Diverse preconditions in the same context (different spans, entity pairs)
- Unaffordable inference cost (since NLI usually uses cross-encoders)

Challenge: Very Large Decision Spaces

Entity Linking, Fine-grained Typing



Extreme multi-label classification (XMLC)


WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes

Article Talk Read Edit View his

Support-vector machine

From Wikipedia, the free encyclopedia

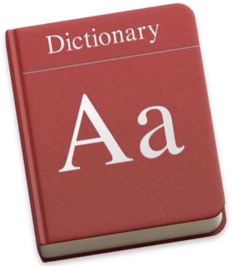
In machine learning, support-vector machines (SVMs, also support-vector networks[1]) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. Developed at AT&T Bell Laboratories by Vladimir Vapnik with colleagues (Boser et al., 1992, Guyon et al., 1993, Cortes and Vapnik, 1995,[1] Vapnik et al., 1997[citation needed]) SVMs are one of the most robust prediction methods, being based on statistical learning frameworks or VC theory proposed by Vapnik (1982, 1995) and Chervonenkis (1974). Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). SVM maps training examples to points in space so as to maximise the width of the gap between the two categories. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

When data are unlabelled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups.

Tags:
machine learning
AI
supervised learning

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



Thousands to millions of labels, more like a dictionary.

Challenge: Non-discriminative Decision Making

Non-discriminative or structured decisions that are beyond the ability of NLI

Spans (Extractive QA)

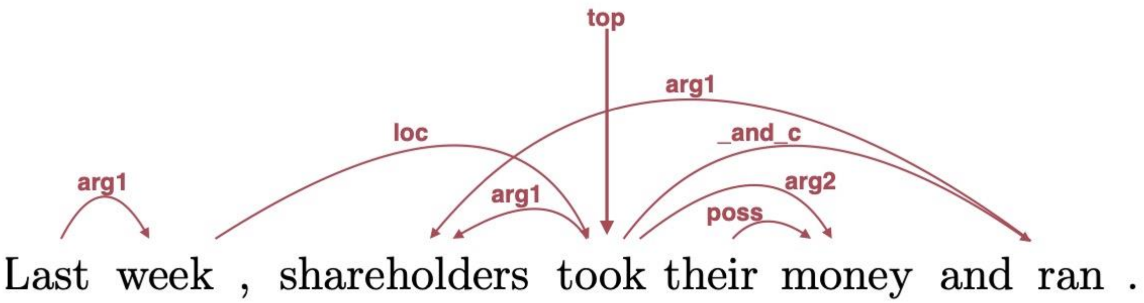
Passage Context

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?

Structures (SDP)



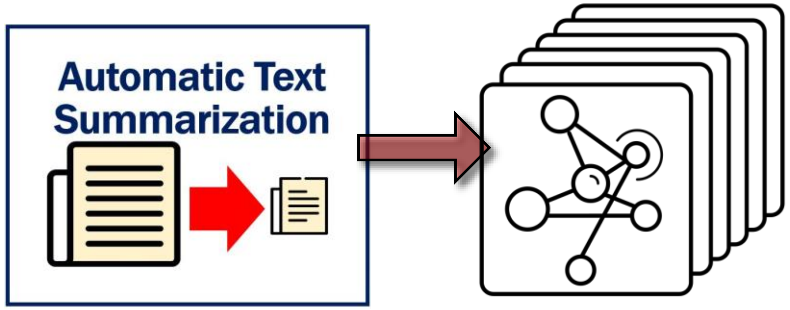
Generation (QFS)

Query: *"Describe the coal mine accidents in China and actions taken"*

Example summary (from Li and Li 2013):
(1) In the first eight months, the death toll of coal mine accidents across China rose 8.5 percent from the same period last year. (2) China will close down a number of ill-operated coal mines at the end of this month, said a work safety official here Monday. (3) Li Yizhong, director of the National Bureau of Production Safety Supervision and Administration, has said the collusion between mine owners and officials is to be condemned. (4) from January to September this year, 4,228 people were killed in 2,337 coal mine accidents. (5) Chen said officials who refused to register their stakes in coal mines within the required time

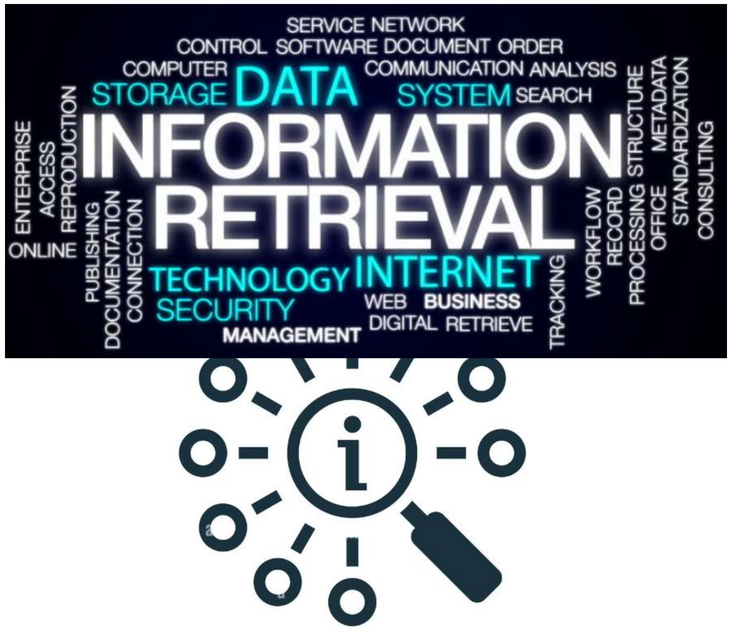
In This Talk

1. Constrained Generation as Indirect Supervision



2. QA as Indirect Supervision

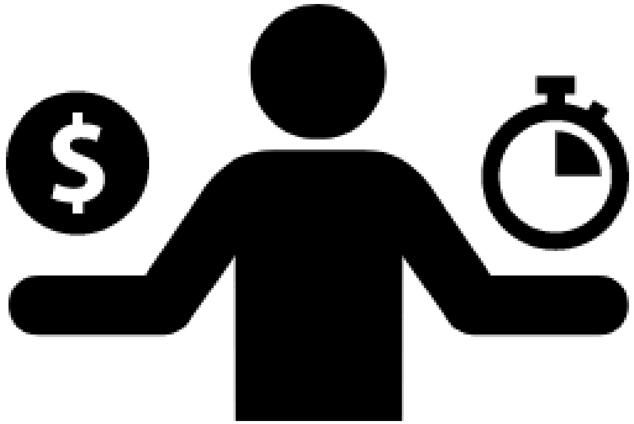
3. IR as Indirect Supervision



Insufficient Structural Annotations

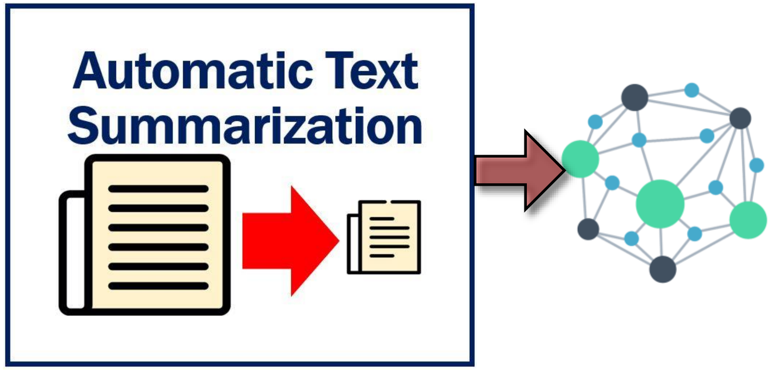
Information extraction suffers from insufficient supervision

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

Head Mention **Tail Mention**

Context (Sentence) Mandelbrot was born in Poland but as a child moved to France

Benoit Mandelbrot

Type: Person

Head Entity

Relation?

- Title
- City of Birth
- Country of Birth

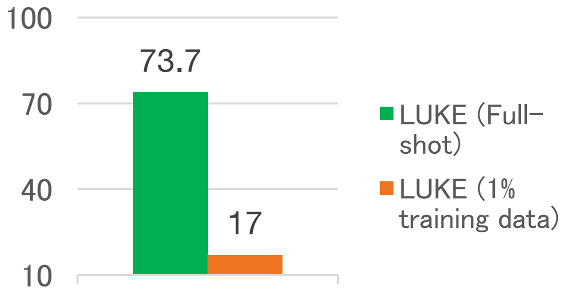
Poland

Type: Country

Tail Entity

Formulated As Multi-class Classification

- Heavily relying on enough relation annotations



The model almost cannot generalize to rarely seen or unseen relations.

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.]

Summarize
(seq2seq gen)

Summary

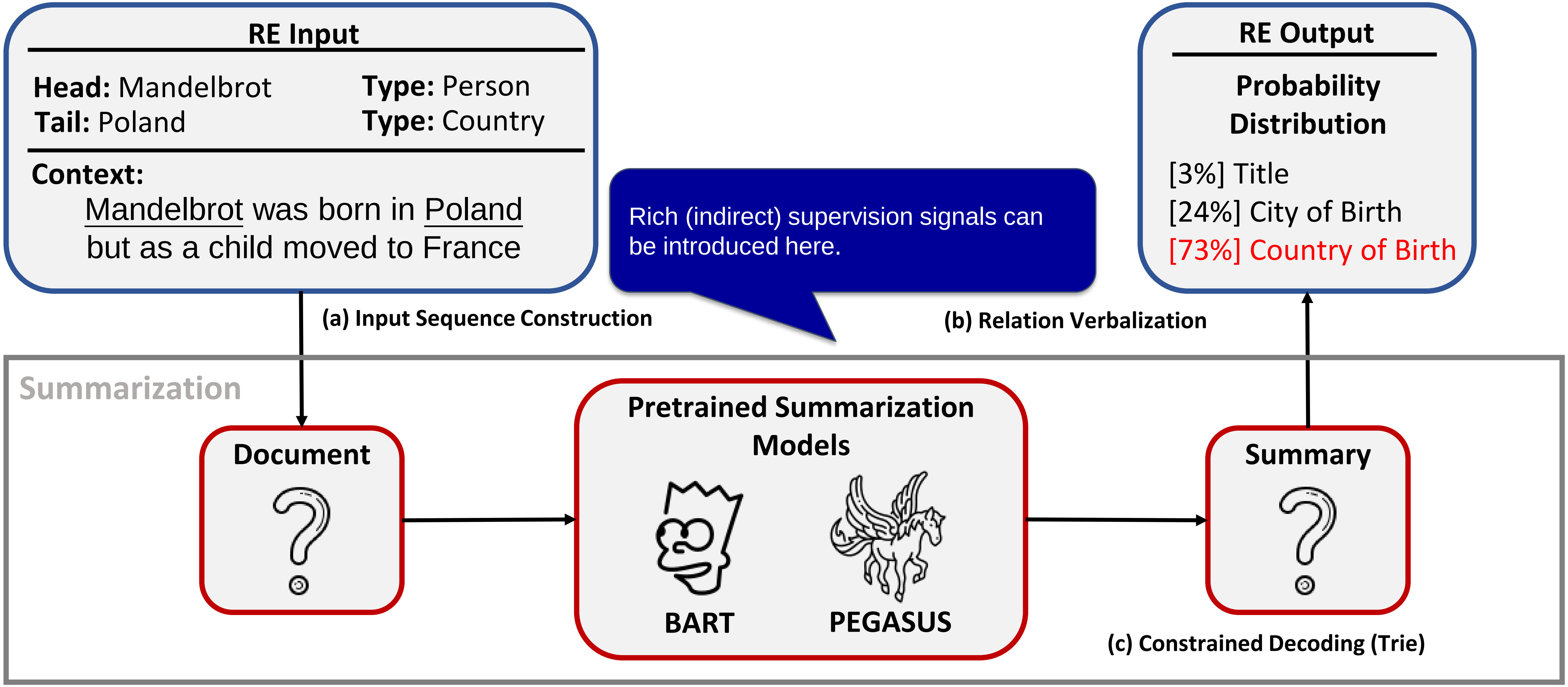
A man and a child have been killed after a light aircraft made an emergency landing on a beach in Portugal.

A more resource-rich task

- Million-scale parallel summary corpora (vs. a few hundred docs or less than 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



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.

Entity Information Verbalization

Input Sequence

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.

Relation Verbalization

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

Both become natural language text that fits a summarization model.

	Relations		Templates
r_1	title	T_1	{subj} is a {obj}
r_2	city of birth	T_2	{subj} was born in the city {obj}
r_3	country of birth	T_3	{subj} was born in the country {obj}
r_4	founded by	T_4	{subj} was founded by {obj}
r_5	NA	T_5	{subj} has no known relations to {obj}

Verbalization →

Relation Verbalization:

r_1: Mandelbrot is a Poland

r_2: Mandelbrot was born in the city Poland

r_3: Mandelbrot was born in the country Poland

r_4: Mandelbrot was founded by Poland

r_5: Mandelbrot has no known relation to Poland

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

?

RE Output

Probability Distribution

[3%] Title
[24%] City of Birth
[73%] Country of Birth

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

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

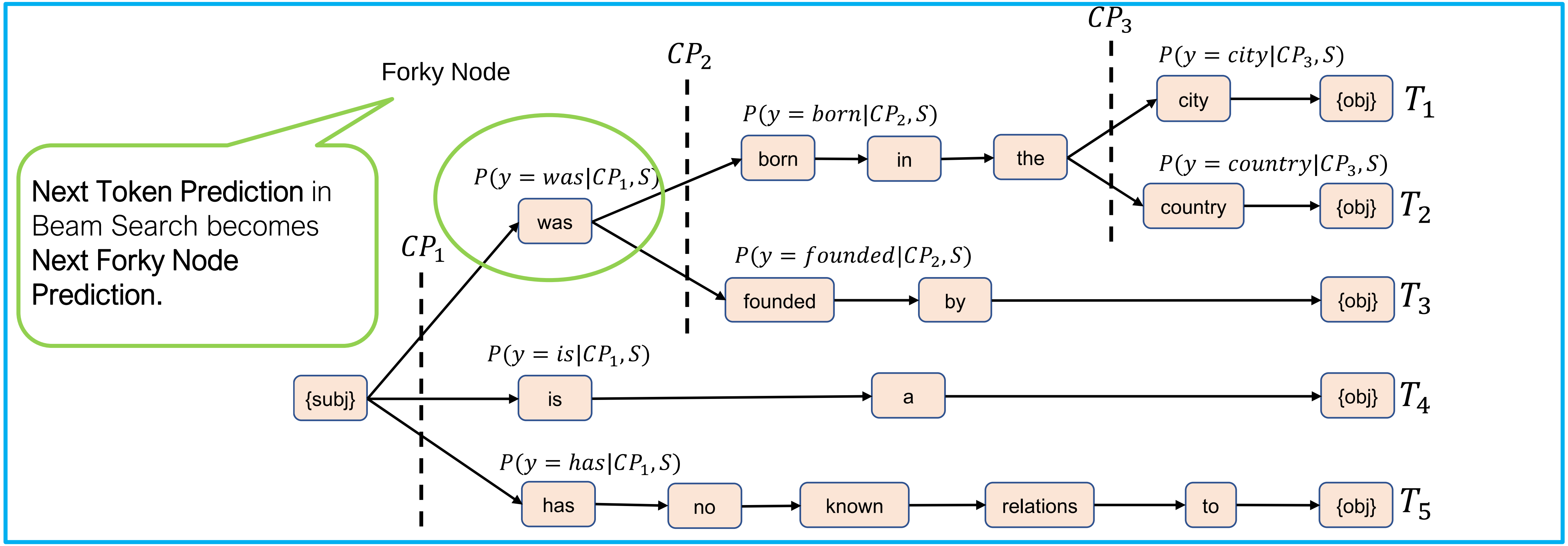
Inference with Trie-based Constrained Decoding

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

Step 1. Build a Trie for relations

Step 2. Beam Search on the Trie

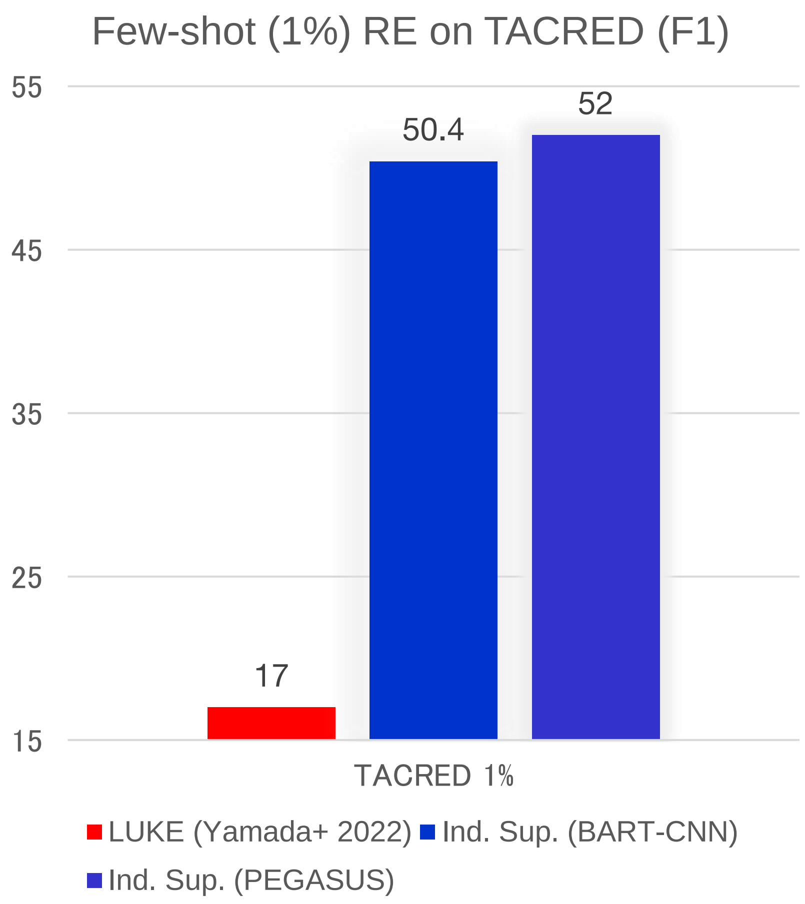
Step 3. Calculate accumulate scores

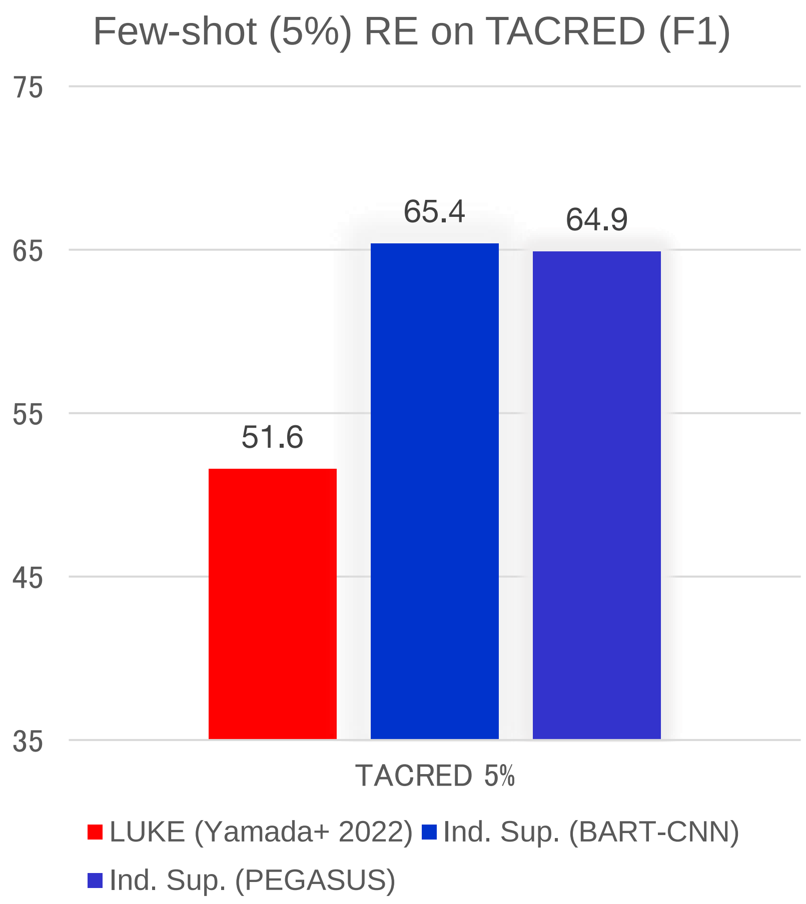


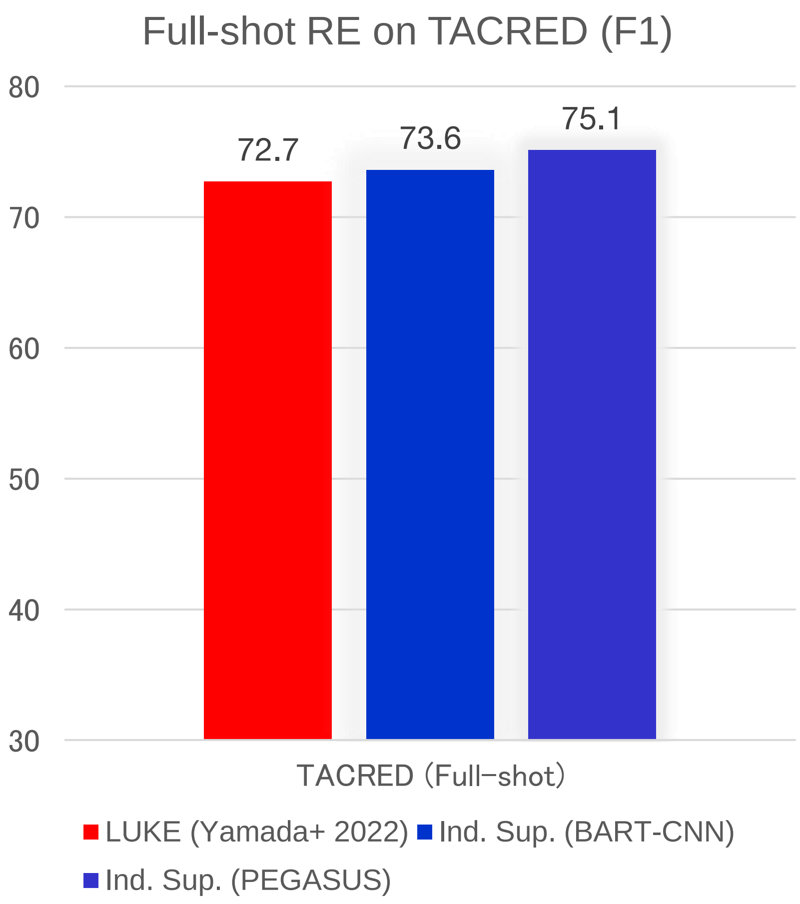
$$P(r_1) = P(y = \text{was}|CP_1, S)P(y = born|CP_2, S)P(y = city|CP_3, S)$$
$$P(r_2) = P(y = \text{was}|CP_1, S)P(y = born|CP_2, S)P(y = country|CP_3, S)$$

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

Summarization Results in Strong Indirect Supervision







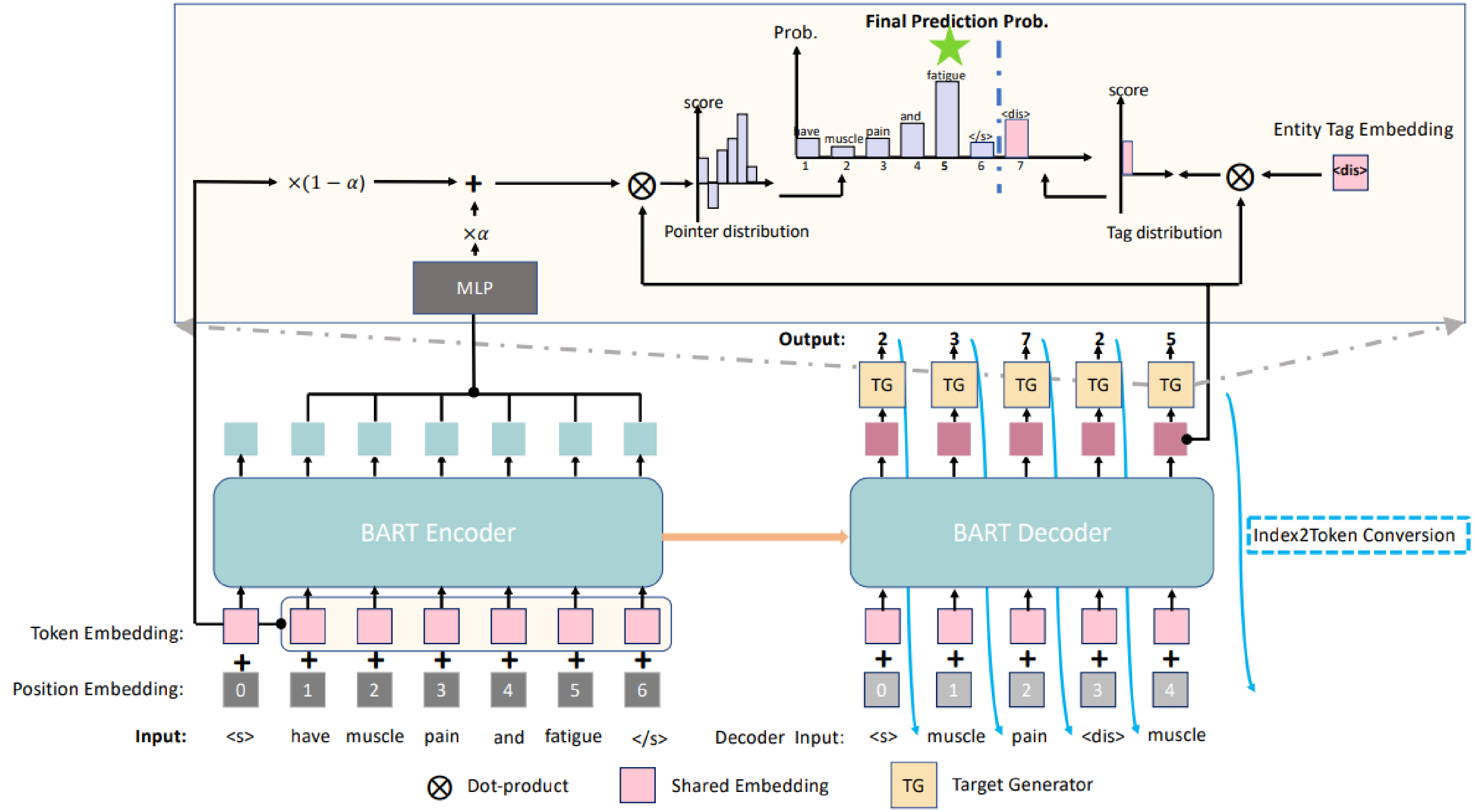
Summarization provides strong indirect supervision for low-resource relation extraction.

Also leading 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

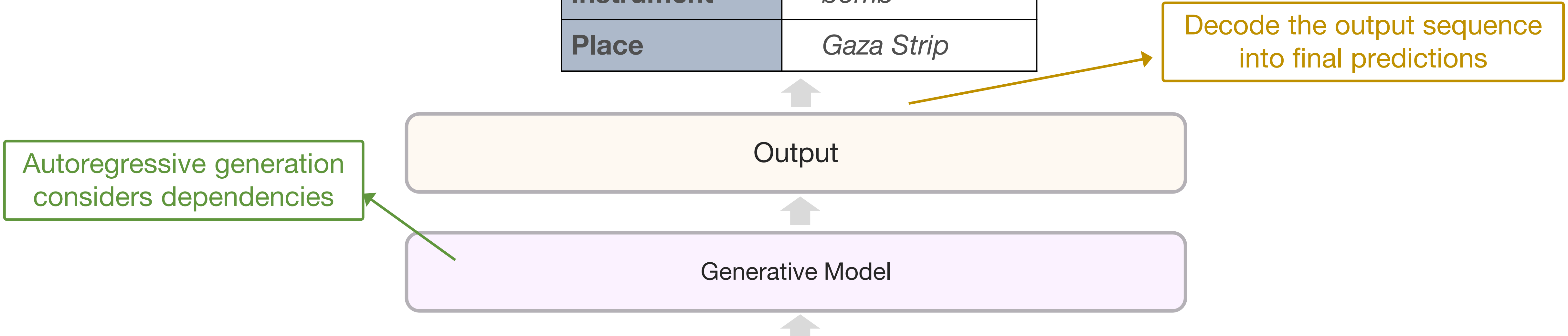


Yan et al. A Unified Generative Framework for Various NER Subtasks. **ACL** 2021

Generative Event Extraction

Event extraction as a conditional generation problem

Event Trigger	*detonated*
Attacker	*Palestinian*
Target	*jeep, soldiers*
Instrument	*bomb*
Place	*Gaza Strip*

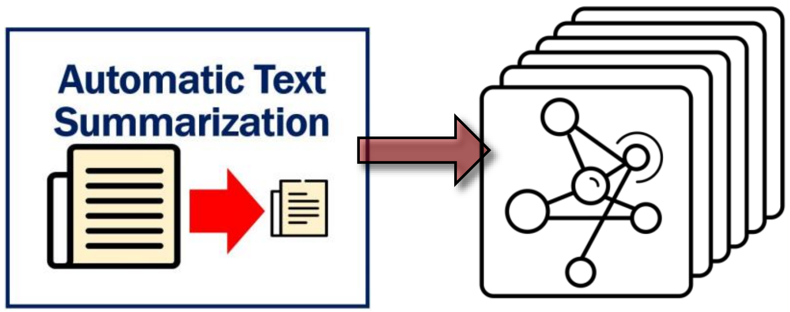


Earlier Monday , a 19-year-old *Palestinian* riding a bicycle *detonated* a 30-kilo (66-pound) *bomb* near a military *jeep* in the *Gaza Strip* , injuring three *soldiers*.

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.)

Span or structural decisions.

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.

Free-form decisions

NER as Extractive QA

Benefit 1: Handling nested entity mentions (not feasible for sequence tagging)

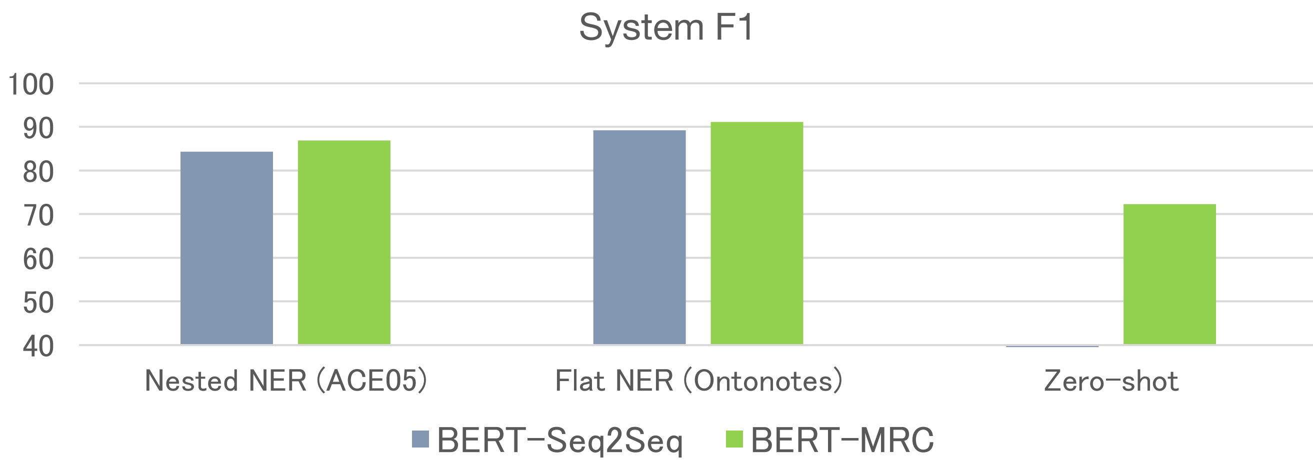
Last night, at ***the Chinese embassy in France***, there was a holiday atmosphere .

Find facilities in the text, including buildings, airports, highways and bridges.

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

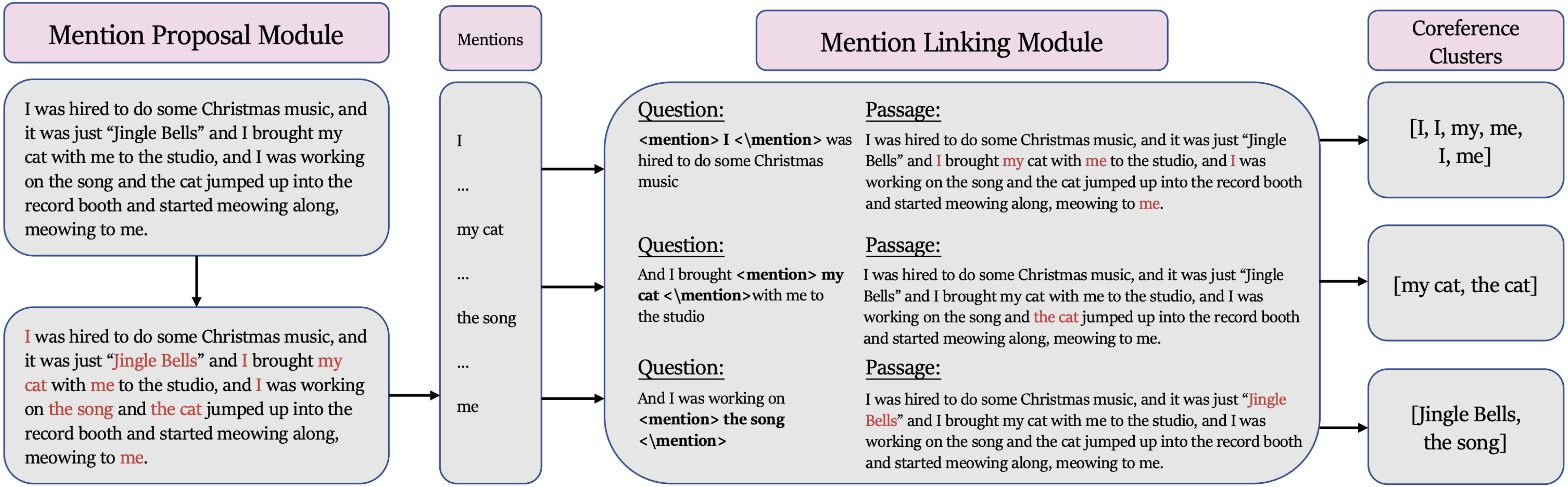
Entity	**Natural Language Question**
Location	Find locations in the text, including non-geographical locations, mountain ranges and bodies of water.
Facility	Find facilities in the text, including buildings, airports, highways and bridges.
Organization	Find organizations in the text, including companies, agencies and institutions.

Better performance than seq2seq generation



Li et al. A Unified MRC Framework for Named Entity Recognition. **ACL** 2020

Coreference Resolution as Extractive QA



Using the sentence that each mention is in as the "question", all other spans belonging to the same cluster as "answers"

Wu et al. CorefQA: Coreference Resolution as Query-based Span Prediction. **ACL** 2020

Other Tasks as Extractive QA

What did someone purchase?
Wade **purchased** Cunningham's home
Who purchased something?
in San Diego for over $1.6M

Federal agencies are investigating Rep.

What did someone sell to someone?
Cunningham for **selling** his house to
Mitchell Wade
Who did someone sell something to?

The plane took off in Los Angeles. The tourists will **arrive** in Mexico at noon.

entity in motion	Who will arrive in Mexico?
end point	Where will the tourists arrive?
start point	Where will the tourists arrive from?
manner	How will the tourists arrive?
cause	Why will the tourists arrive?
temporal	When will the tourists arrive?

QA-SRL: QA as Semantic Role Labeling

Relation	Question	Sentence & Answers
educated_at	What is **Albert Einstein**'s alma mater?	**Albert Einstein** was awarded a PhD by the **University of Zürich**, with his dissertation titled...
occupation	What did **Steve Jobs** do for a living?	**Steve Jobs** was an American **businessman**, **inventor**, and **industrial designer**.
spouse	Who is **Angela Merkel** married to?	**Angela Merkel**'s second and current husband is quantum chemist and professor **Joachim Sauer**, who has largely...

QA for Relation Extraction

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 a. Large-Scale QA-SRL Parsing. ACL 2018
Levy et al. Zero-Shot Relation Extraction via Reading Comprehension. CoNLL 2017

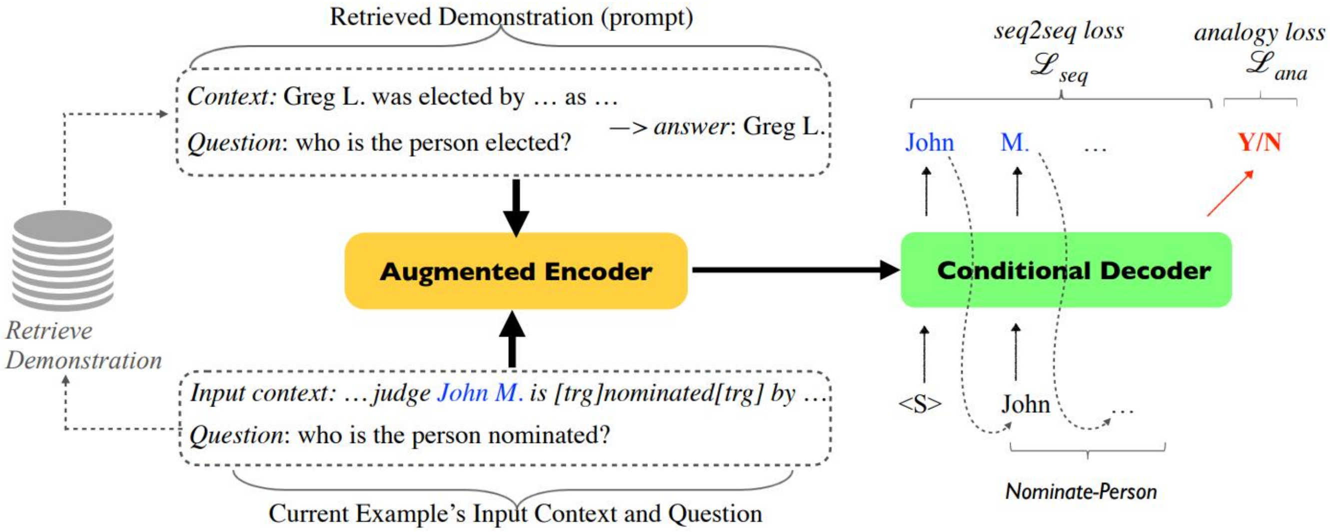
Tasks Reformulated as Abstractive QA

Benefits of An Abstractive QA Reformulation

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

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.

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



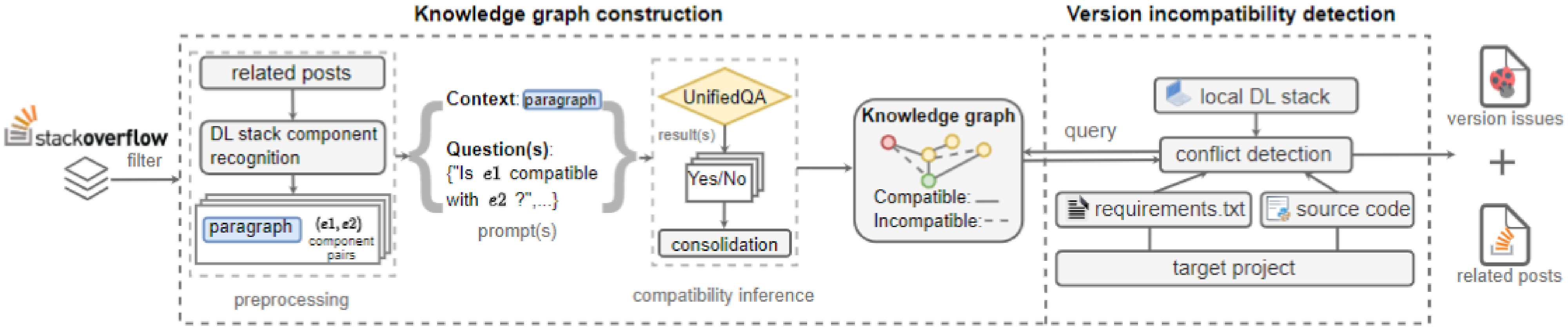
Event argument generation

Zhao et al. Compositional Task-Oriented Parsing as Abstractive Question Answering. NAACL 2022
Du and Ji. Retrieval-Augmented Generative Question Answering for Event Argument Extraction. EMNLP 2022

Specialized Domain Application

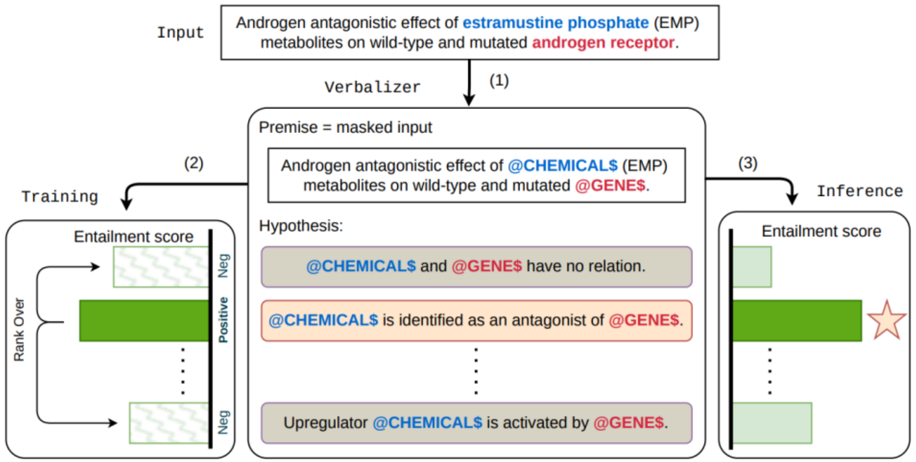
Generalizability and lack of annotations are more significant challenges here

QA for software version compatibility detection



Zhao et al. Knowledge-based Version Incompatibility Detection for Deep Learning. **ESEC/FSE** 2023

Extracting drug-drug interaction



Clinical event extraction

Sign_symptom

A man presented with an abnormal nodule measuring 0.8 x 1.5 cm in the left upper lung lobe imaged through chest computed tomography scanning.

Diagnostic_procesure

Event trigger	nodule
Event type	Sign_symptom
Detailed description	abnormal
Area	0.8 x 1.5 cm
Biological structure	left upper lung lobe

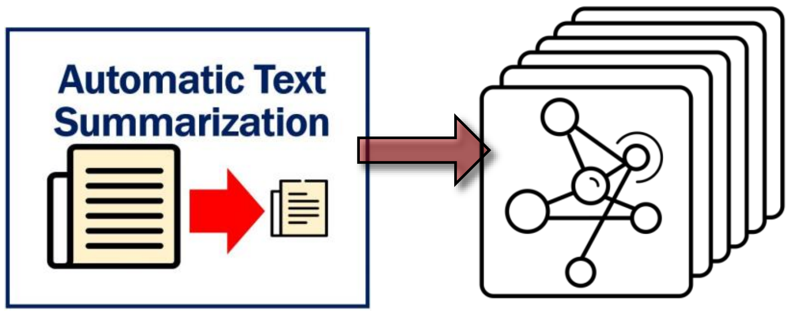
Event trigger	computed tomography
Event type	Diagnostic_ procedure
Biological structure	chest

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

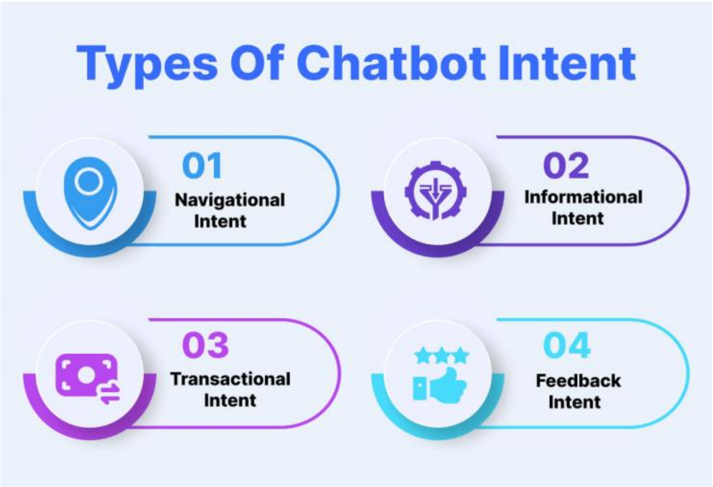


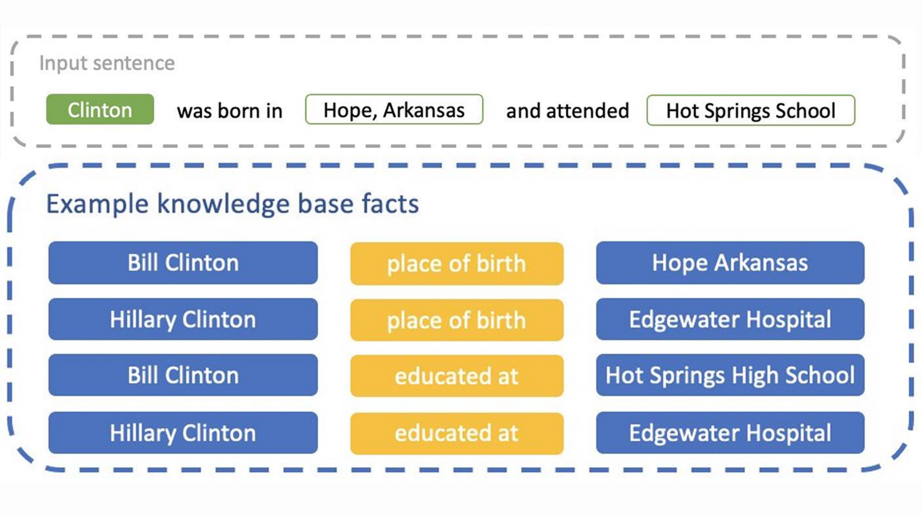
2. QA as Indirect Supervision

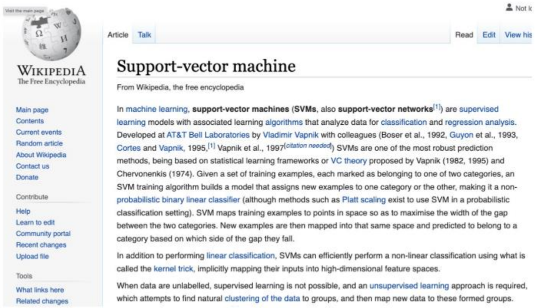
3. IR as Indirect Supervision

Dense Retrieval for NLU Tasks with Large Decision Spaces

Some NLU tasks may have very large decision spaces







Tags:
machine learning
AI
supervised learning

Intent Detection

- Target: Hundreds of intent types

Entity Typing and Linking

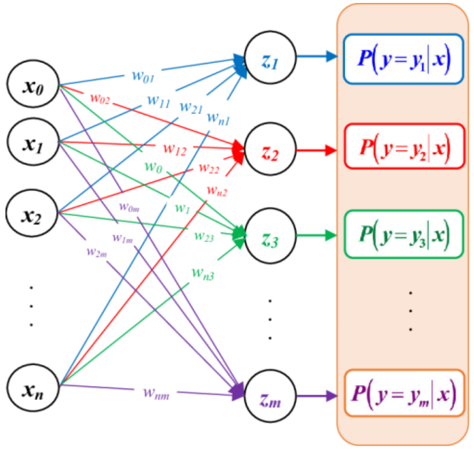
- Target: entities in a whole KB

Extreme multi-label classification (XMLC)

- Thousands to millions of tags

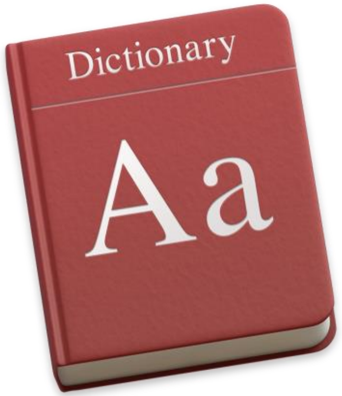
Decision spaces in a hundred- to million-scale

Dense Retrieval for NLU Tasks with Large Decision Spaces



Supervising a classifier is not ideal

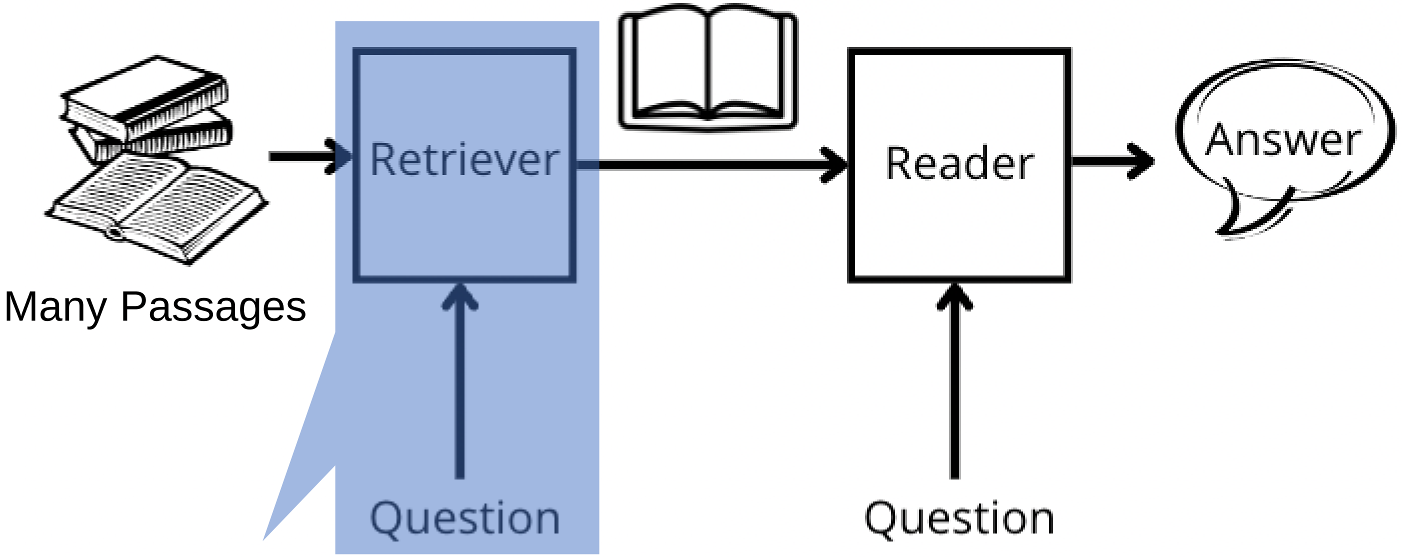
- Too few instances per class
- Meaningless class label representation
- Not generalizable to unseen classes



Learning to lookup a label thesaurus should be more feasible

- A plausible source of indirection supervision: Dense Retrievers
- Meaningful label representation
- Retrieval generalizes to unseen labels

Dense Retriever for Passage Retrieval in Open-domain QA

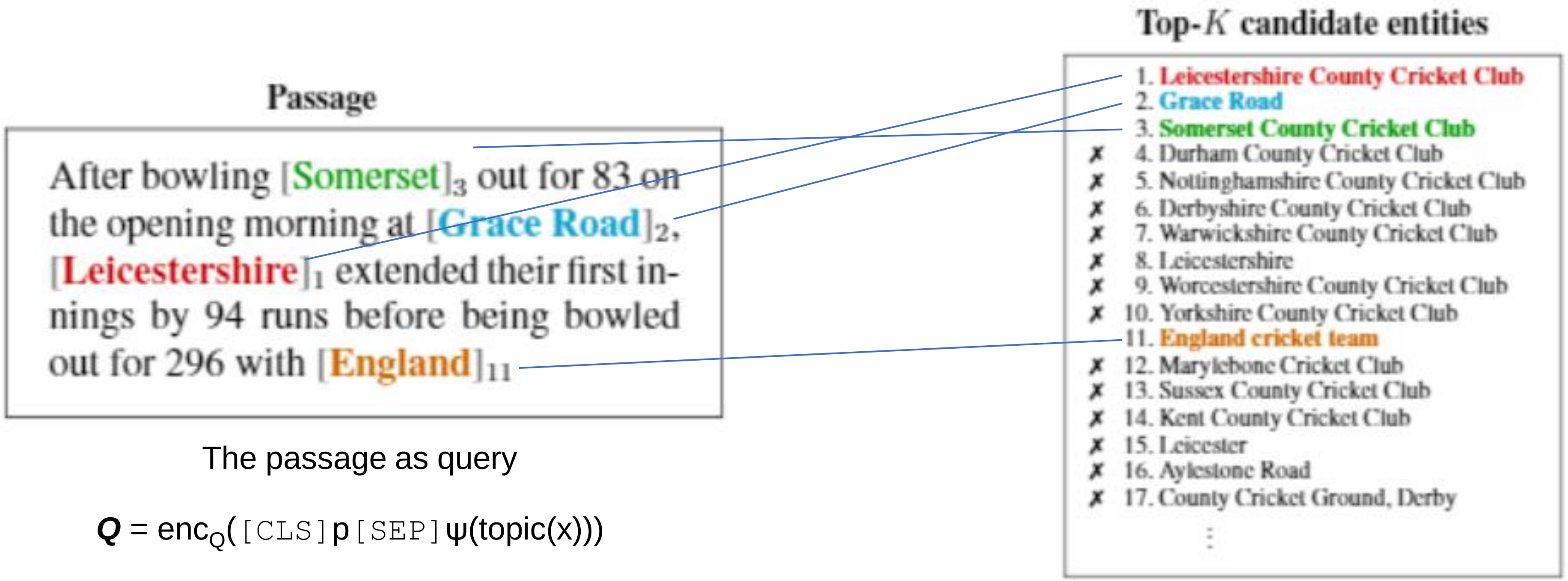


A dual-encoder model for retrieving passages most relevant to a question

- Two encoders $\boldsymbol{P}$ and $\boldsymbol{Q}$ for passages and questions
- Contrastive learning to maximize $\boldsymbol{P}^T \cdot \boldsymbol{Q}$ for correct question-passage pair
- Efficient and generalizable retrieval (using MIPS)

Karpukhin et al. Dense passage retrieval for open-domain question answering. EMNLP 2020

Dense Retrieval for Entity Linking



The passage as query

$\boldsymbol{Q} = \mathrm{enc}_Q(\texttt{[CLS]}\, p\, \texttt{[SEP]}\, \psi(\mathrm{topic}(x)))$

Entity candidates with descriptions

$\boldsymbol{P} = \mathrm{enc}_P(\texttt{[CLS]}\, \phi\mathrm{title}(e) \oplus \phi\mathrm{desc}(e)\, \texttt{[SEP]})$

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

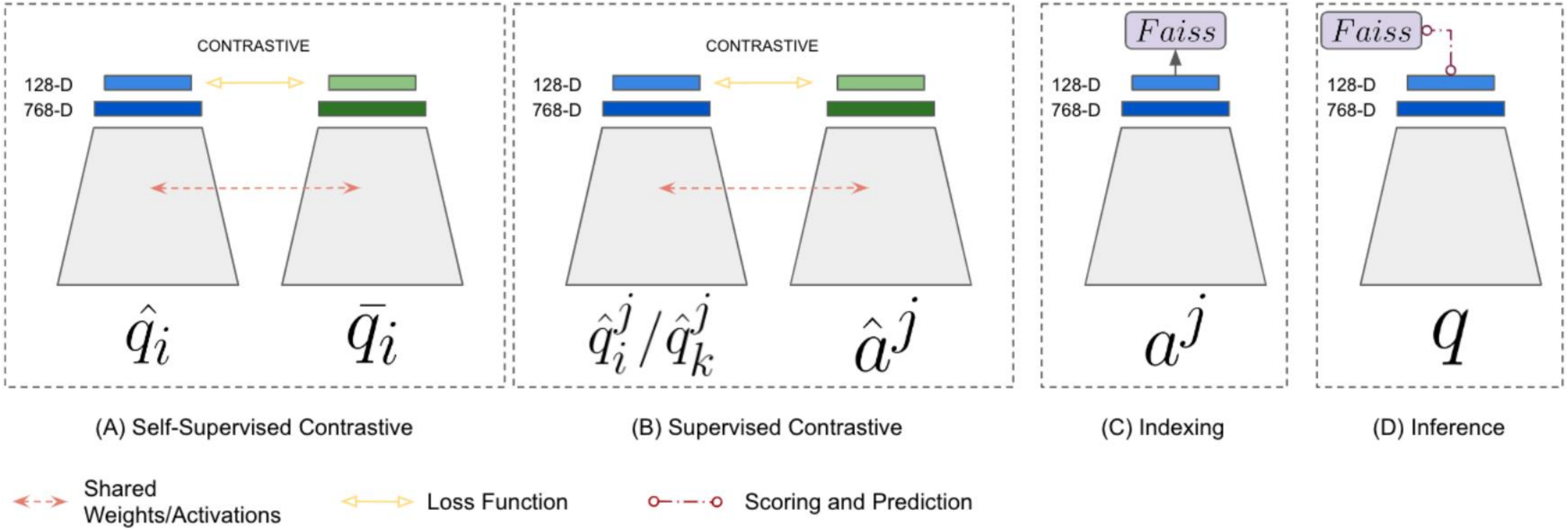
Zhang et al. EntQA: Entity Linking as Question Answering. ICLR 2022

Dense Retrieval for (Few-shot) Intent Prediction

"How long will it take for me to get my card?"
"Can you tell me how long it takes for a new card to come?"
"Can you tell me the status of my new card?"
"how many days processing new card?"

→

`card_arrival`
`card_delivery_estimate`
`lost_or_stolen_card`
`contactless_not_working`

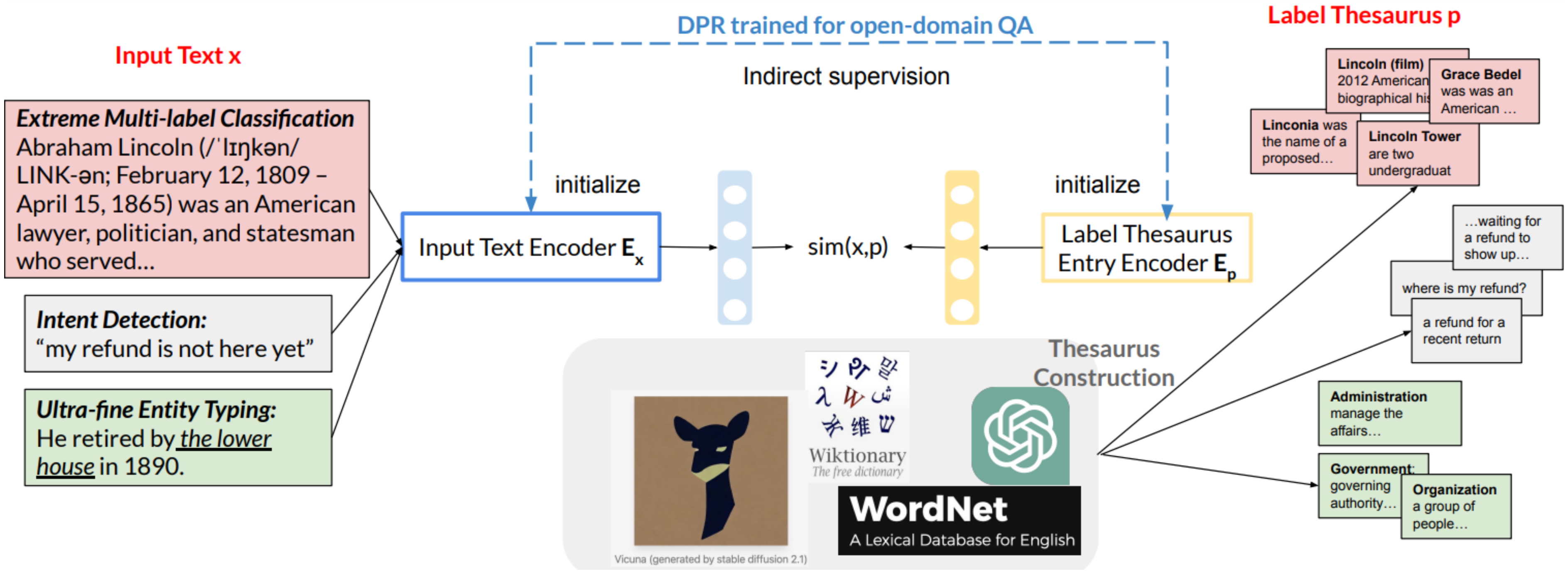


Indirect supervision for retrieving from a fine-grained pool of intents

Enhancing few-shot generalizability (+5.22~8.50% in accuracy for 5-shot prediction)

Yehudai et al. QAID: Question Answering Inspired Few-shot Intent Detection. ICLR 2023

Dense Retrieval as A General Solution



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

Xu et al. Dense Retrieval as Indirect Supervision for Large-space Decision Making. 2023

Dense Retrieval as A General Solution

Indirect Supervision Improves Three Large-space Decision Making Tasks

Input Text


Abraham Lincoln
193 languages
Article Talk Tools
From Wikipedia, the free encyclopedia

This article is about the president of the United States. For other uses, see Abraham Lincoln (disambiguation).

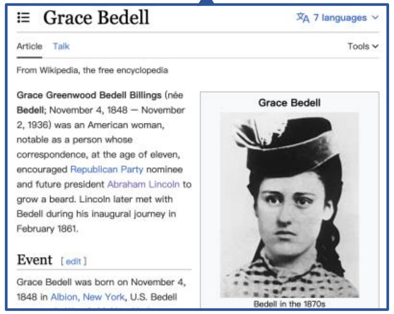
Abraham Lincoln (/ˈlɪŋkən/ LINK-ən; February 12, 1809 – April 15, 1865) was an American lawyer, politician, and statesman who served as the 16th president of the United States from 1861 until his assassination in 1865. Lincoln led the Union through the American Civil War to defend the nation as a constitutional union and succeeded in abolishing slavery, bolstering the federal government, and modernizing the U.S. economy.

Lincoln was born into poverty in a log cabin in Kentucky and was raised on the frontier, primarily in Indiana. He was

Abraham Lincoln

Portrait by Alexander Gardner, 1863

Label Thesaurus Entry


Grace Bedell
7 languages
Article Talk Tools
From Wikipedia, the free encyclopedia

Grace Greenwood Bedell Billings (née **Bedell**; November 4, 1848 – November 2, 1936) was an American woman, notable as a person whose correspondence, at the age of eleven, encouraged Republican Party nominee and future president Abraham Lincoln to grow a beard. Lincoln later met with Bedell during his inaugural journey in February 1861.

Event [edit]

Grace Bedell was born on November 4, 1848 in Albion, New York, U.S. Bedell

Grace Bedell

Bedell in the 1870s


The Towers (Ohio State)
Add languages
Article Talk Tools
From Wikipedia, the free encyclopedia

Abraham Lincoln Tower & Justin S. Morrill Tower, also known as **The Towers**, **Morrill Tower** or **Lincoln Tower** are two undergraduate residential houses at The Ohio State University. The Towers are located on the west campus of the Ohio State University across from the Drake Union off the east banks of the Olentangy River. Morrill Tower is to the right of Ohio Stadium on Cannon Drive. The towers are in close proximity of OSU's RPAC (Recreation and Physical Activity Center) and the Wexner Medical Center.

Justin S. Morrill Tower & Abraham Lincoln Tower

The Towers

Lincoln *(front)* & Morrill *(back)* Towers


Lincoln (film)
50 languages
Article Talk Tools
From Wikipedia, the free encyclopedia

Lincoln is a 2012 American biographical historical drama film directed and produced by Steven Spielberg, starring Daniel Day-Lewis as United States President Abraham Lincoln.[8] It also features Sally Field, David Strathairn, Joseph Gordon-Levitt, James Spader, Hal Holbrook and Tommy Lee Jones in supporting roles.

The screenplay by Tony Kushner was loosely based on Doris Kearns Goodwin's 2005 biography *Team of Rivals: The Political Genius of Abraham Lincoln*, and covers the final four

Lincoln

① Extreme Multi-label Classification (XMC)

Method	Precision @1	Precision @3	Precision @5	Recall @1	Recall @3	Recall @5
MACLR	20.99	15.57	12.26	12.59	23.94	29.41
DDR	**27.78**	**19.98**	**15.66**	**15.03**	**28.87**	**35.47**

② Ultra-fine Semantic Typing

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 .

Label Thesaurus Entry

politician
a leader engaged in civil administration

person
a human being

campaigner
a politician who is running for public office

Method	Precision	Recall	F1
Context-TE	53.7	49.4	51.5
DDR	51.9	**52.3**	**52.1**

+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

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

Thank You