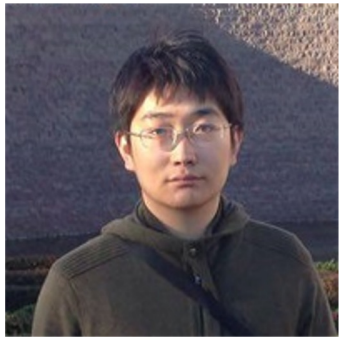
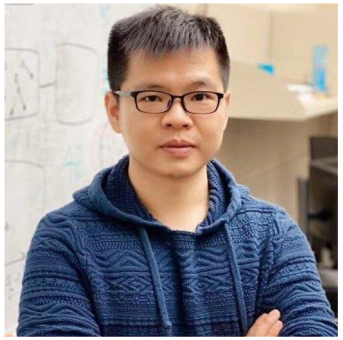
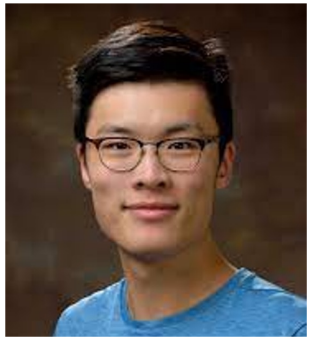
New Frontiers of Information Extraction


Muhao Chen


Lifu Huang


Manling Li


Ben Zhou


Heng Ji


Dan Roth

July 2022

NAACL Tutorial

New Frontiers of Information Extraction

NAACL 2022

Information Extraction

- Understanding text depends on the ability to Extract Information from it
  - Identify and contextualize
    - entities,
    - quantities (and their scope),
    - events,
    - relations, etc.
  - Often, there is a need to disambiguate and link entities and events to encyclopedic resources

In the first quarter, the Bears Cutler fired a 7-yard TD pass to tight end Greg Olsen. … In the third quarter, the … back Adrian Peterson's 1-yard touchdown run. The Bears increased their lead over the Vikings with Cutler's 2-yard TD ss to tight end Desmond Clark. The gap was reduced when Favre fired -yard TD pass to tight end Visanthe Shiancoe. The Vikings … with ian Peterson's second 1-yard TD run. The Bears then responded w Cutler firing a 20-yard TD pass to wide receiver Earl Bennett. T ears then won on Jay Cutler's game-winning 39-yard TD pass to receiver Devin Aromashodu.

She orts worsened seizure frequency, seizures now occurring u 10/week, in clusters about 2-3 day/week. Previously orted seizures occurring about 2-3 times per month, often round the time of menses,…

Mayor Rahm Emanuel has raised more than $10 million toward his bid for a third term – more than five times the total raised by his 10 challengers combined, campaign finance records show.

The COVID-19 pandemic in the United States is part of the worldwide pandemic of coronavirus disease 2019 (COVID-19). As of October 2020, there were more than 9,000,000 cases and 230,000 COVID-19-related deaths in the U.S., representing 20% of the world's known COVID-19 deaths, and the most deaths of any country.

Information Extraction

- Understanding text depends on the ability to Extract Information from it
  - Identify and contextualize
    - events,
    - entities,
    - quantities (and their scope),
    - relations, etc.
  - Often, there is a need to disambiguate and link entities and events to encyclopedic resources
- Why do we need to extract information?
  - Facilitate answering questions about the text
- IE is the backbone of any knowledge-driven AI system
  - Needed even for evaluation, summarization

Who scored the longest touchdown pass of the game?

In the … Greg … touchdown run. The Bears increased their lead over the Vikings with Cutler's 2-yard TD pass to tight end Desmond Clark. The gap was reduced when Favre fired a 6-yard TD pass to tight end Visanthe Shiancoe. The Vikings … with Adrian Peterson's second 1-yard TD run. The Bears then responded with Cutler firing a 20-yard TD pass to wide receiver Earl Bennett. The Bears then won on Jay Cutler's game-winning 39-yard TD pass to wide receiver Devin Aromashodu.

What is her seizure frequency?

She reports wors… now occurring up to 10/week, in clusters about 2-3 day/week. Previously reported seizures occurring about 2-3 times per month, often around the time of menses,…

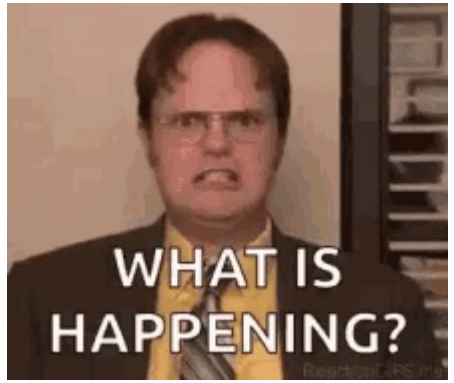
How much did his challengers raise?

Mayor Rahm Em… toward his bid for a thi… raised by his 10 challengers combined, campaign finance records show.

The COVID-19 pandemic in the United States is part of the worldwide pandemic of coronavirus disease 2019 (COVID-19). As of October 2020, there were more than 9,000,000 cases and 230,000 COVID-19-related deaths in the U.S., representing 20% of the world's known COVID-19 deaths, and the most deaths of any country.

The Role of Information Extraction

- Most of the work in NLP focuses on understanding "what the text says"
  - We analyze "what's written here" at the sentence level (mostly)
    - But more and more also at a document level
    - And even, at a multi-document level



- Integrating multiple IE tasks allow us to go beyond "what the text says"
  - We can now attend to "what is happening"
    - And ground it in the world

NexLP - Baseball Card | NexLP - Email Thread Vie

localhost/nexlp/threadnav/index

NexLP View Original

Doc: 1 of 1 CASE_000001

James Haggins 6/30/2013 7:28:00 AM
To: James Haggins; James Haggins; Laura Haggins
Subject:Re: rumors
f765488113c744f5a8510911db140ab0

mom,

take a look at the lease I sent to you last night, should be final.

Can you give a shout on the mobile to Jack and Steve? I gave you their numbers. See if they are willing to come with me. Keep it on the down low.

Love you,

Jim

Laura Higgins 7/1/2013 8:03:00 AM
To: James Haggins
Subject:Re: rumor
44ed3bf6ea804fea9bab0b9e47063d17

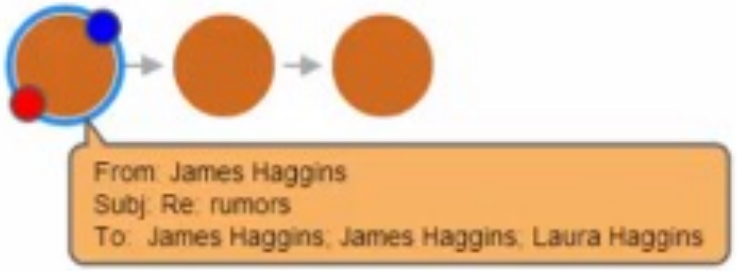
Didn't you want me to take a look at the lease?

Doc: 1 of 1 CASE_000001

Threads | Inclusive Docs | Included Docs | Recommended

Subject	Start	Hits	Doc Count	Seg
Re: rumor	6/29/2013 11:28:(	1	4	
Re: Next visit to London	6/27/2013 11:12:(	6	35	

25 | Page 1 of 1 | 1 to 2 of 2



NLP | NEXT GENERATION LANGUAGE PROCESSING

Date: Thu, 11 May 2000 07:32:00 -0700 (PDT)
From: david.delainey@enron.com
To: mike.jakubik@enron.com
Subject: Re: Raptor

That may work - I don't want to end up with an equity position I just worked hard to eliminate. Further, raptor may be a good accounting hedge but if we took back JEDI's share at existing marks we would be destroying significant real value. Will the buy back price of the equity/debt we get back from JEDI incorporate the write downs we think should occur?

Regards,
Delainey

NLP | NEXT GENERATION LANGUAGE PROCESSING

SENT

Year/Month: May 2000
Month: May
Year: 2000
Work Shift: Business Evening

COMMUNICATORS

david delainey
david delainey
delainey david w
dave delainey
david delaney
dave delaney
david delaine
Mike Jakubik

ORGANIZATIONS

JEDI
JEDI II
Joint Energy Development
Investments

DOMAINS INVOLVED

Enron.com

Date: Thu, 11 May 2000 07:32:00 -0700 (PDT)
From: david.delainey@enron.com
To: mike.jakubik@enron.com
Subject: Re: Raptor

That may work - I don't want to end up with an equity position I just worked hard to eliminate. Further, raptor may be a good accounting hedge but if we took back JEDI's share at existing marks we would be destroying significant real value. Will the buy back price of the equity/debt we get back from JEDI incorporate the write downs we think should occur?

Regards,

Delainey

EMOTIONAL INTELLIGENCE

Sentiment: -5.50
Pressure: 10.50
Opportunity: 7.50
Intent: 4.50
Rationalization: 4.50

CONCEPT AI

Equity
equity interest
equity markets
equity stakes

Hedge
hedge fund
hedging strategy
statistical hedging

Destroying
destroying competition
Destroying value

Raptor
Raptor structure
Raptor vehicles
Raptor transaction

CLUSTER

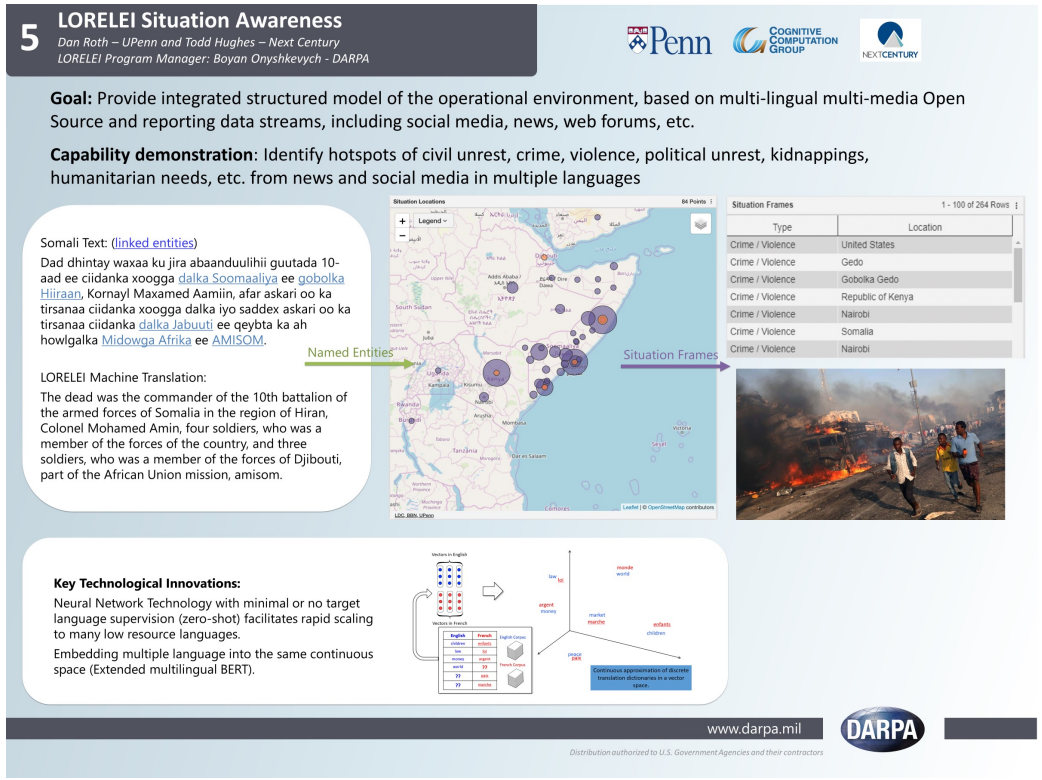
056_electricity/issues/
capacity

Disaster Relief (DARPA-LORELEI Project)

- Information Extraction in Low Resource Language
- Provide integrated model of the operational environment based on streaming data from multiple sources (news, social media, images)

Somali streaming data → Situation Awareness (described in English)

- What is it about?
  - Topics; Events
- "Understand" a situation described in Target Language
  - Identify Entities & Concepts (NER)
  - Ground in English Resources (EDL)



5 LORELEI Situation Awareness

Dan Roth – UPenn and Todd Hughes – Next Century
LORELEI Program Manager: Boyan Onyshkevych - DARPA

Goal: Provide integrated structured model of the operational environment, based on multi-lingual multi-media Open Source and reporting data streams, including social media, news, web forums, etc.

Capability demonstration: Identify hotspots of civil unrest, crime, violence, political unrest, kidnappings, humanitarian needs, etc. from news and social media in multiple languages

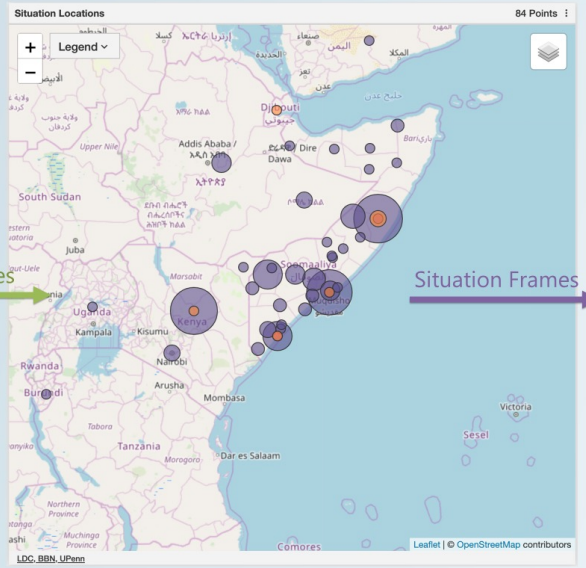
Somali Text: (linked entities)

Dad dhintay waxaa ku jira abaanduulihii guutada 10-aad ee ciidanka xoogga dalka Soomaaliya ee gobolka Hiiraan, Kornayl Maxamed Aamiin, afar askari oo ka tirsanaa ciidanka xoogga dalka iyo saddex askari oo ka tirsanaa ciidanka dalka Jabuuti ee qeybta ka ah howlgalka Midowga Afrika ee AMISOM.

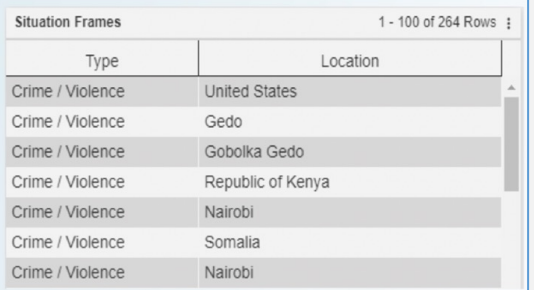
Named Entities

LORELEI Machine Translation:

The dead was the commander of the 10th battalion of the armed forces of Somalia in the region of Hiran, Colonel Mohamed Amin, four soldiers, who was a member of the forces of the country, and three soldiers, who was a member of the forces of Djibouti, part of the African Union mission, amisom.



Situation Frames



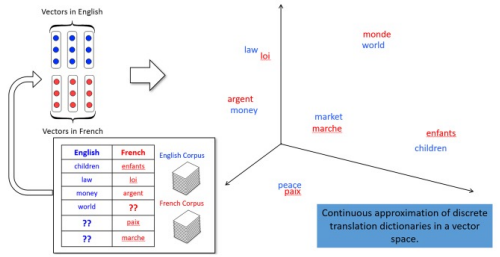
Situation Frames 1 - 100 of 264 Rows

Type	Location
Crime / Violence	United States
Crime / Violence	Gedo
Crime / Violence	Gobolka Gedo
Crime / Violence	Republic of Kenya
Crime / Violence	Nairobi
Crime / Violence	Somalia
Crime / Violence	Nairobi

Key Technological Innovations:

Neural Network Technology with minimal or no target language supervision (zero-shot) facilitates rapid scaling to many low resource languages.

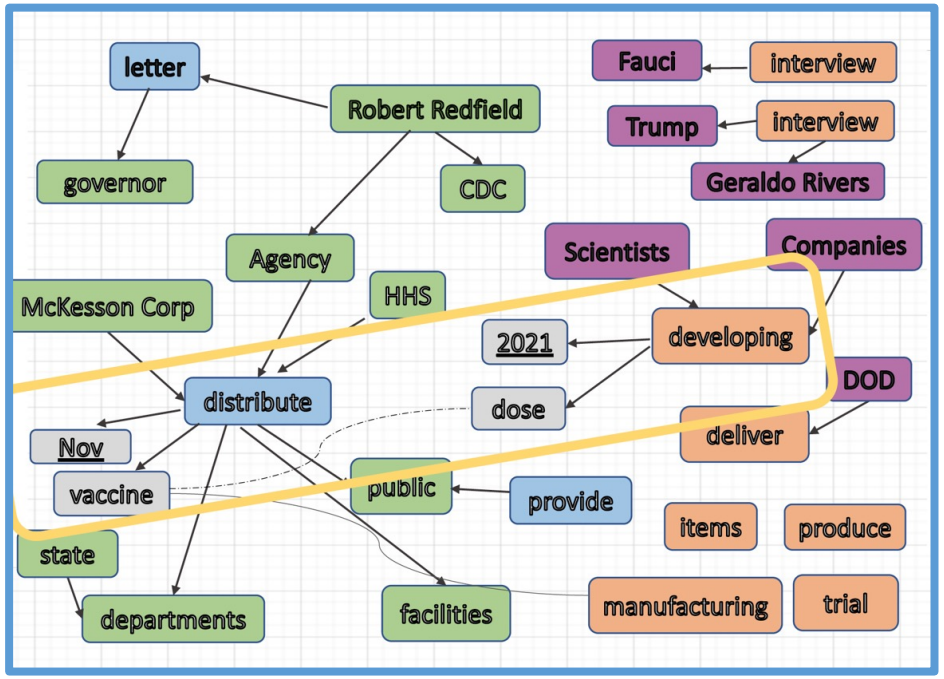
Embedding multiple language into the same continuous space (Extended multilingual BERT).



- We know how to develop models when given a lot of good annotated data
- In many situations, exhaustive annotation is not realistic.
- **Setting:**
  - Two surprise languages
    - 2019: Odia, Ilocano
  - Develop solutions for several IE tasks (in a week)
  - **No annotated data**
  - Some target language data; a (limited quality) dictionary.
  - Minimal remote exposure to native speakers

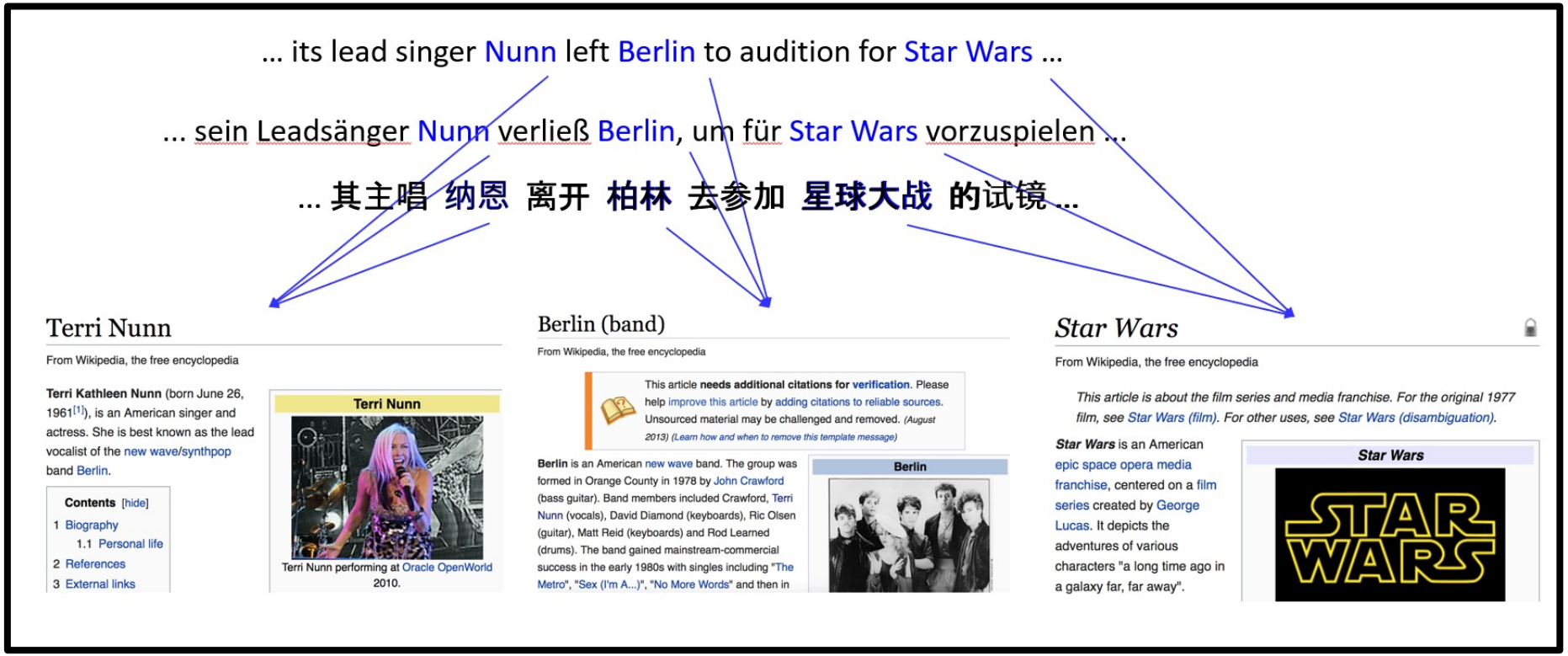
Challenges

- It is not sufficient to do information Extraction at the sentence level
- Often, not even the document level, but rather at the document collection level



Challenges

- Information Extraction cannot be done at a sentence level
- Often, not even at the document level, but rather at the document collection level
- It needs to be multilingual



Challenges

- Information Extraction cannot be done at a sentence level
- Often, not even at the document level, but rather at the document collection level
- It needs to be multilingual
- It needs to be multimodal
  - ☐ News, Encyclopedic sources, Social Media
  - ☐ Images and Videos

- People
- Text
- Masks
- Clothing
- Digital Screens
- …

Fu et. al. ACL'22

When and Where?

Challenges

- Information Extraction cannot be done at a sentence level
- Often, not even at the document level, but rather at the document collection level
- It needs to be multilingual
- It needs to be multimodal
- It needs to be robust
  - Domain, style, …..

Financial domain: Rate Fixing

- Inter-banking rate issues
  - A client investigates if bankers were fixing the inter-banking rate.
  - *"Mate, can you raise the main one by 0.2 till Wednesday? I owe you a drink."*
- Messages here are typically very short; the language is colloquial and ungrammatical. Messages include many quantities (rates), how much to adjust, etc.

Challenges

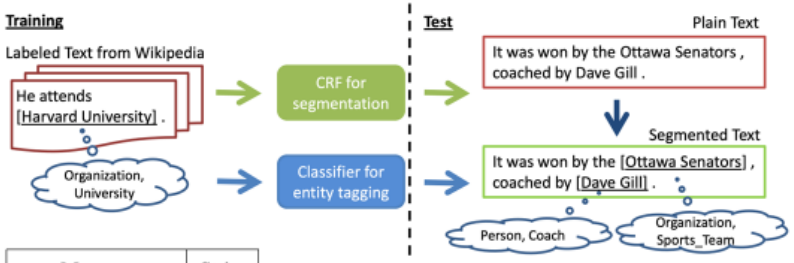
- Information Extraction cannot be done at a sentence level
- Often, not even at the document level, but rather at the document collection level
- It needs to be multilingual
- It needs to be multimodal
- It needs to be robust
- Supervision
  - Too many (ill-defined) decisions
    - Annotating text for all is not scalable
  - Level of granularity is a challenge
  - End-task supervision by itself is often too loose
  - There is a need to resort to exploiting incidental supervision signals [Roth AAAI'17]
    - Self-Supervision from grounding in existing resources

This Tutorial

Minimally and Indirectly Supervised IE

Weak Supervision – Knowledge Bases

- One of the earliest attempts: entity and entity relations
- Ling and Weld (2012): NER from KB supervision



Measure	Strict
NEL	0.220
Stanford (CoNLL)	0.425
FIGER	0.471
FIGER (GOLD)	0.532

Note: StanfordNER only predicts a subset of the taxonomy

Ling and Weld. Fine-Grained Entity Recognition. AAAI 2012.

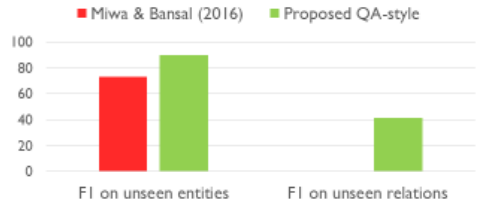
Indirect Supervision from QA

- Levy et al. (2017): Relation extraction formulated as QA

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...

Schema Querification (crowdsourced)

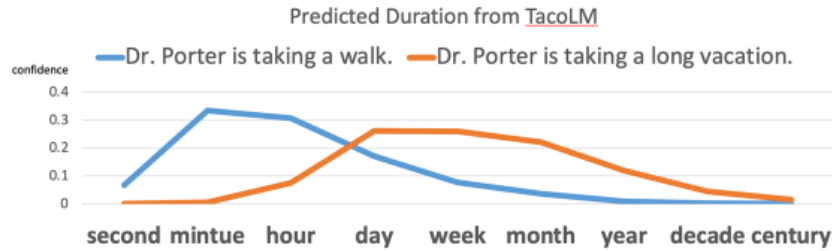
- Why would it work?
  - Question provides "indirect" information on relation labels



41

Weak Supervision – Linguistic Patterns

- Zhou et al. (2020): Temporal Information Extraction from Patterns
- Goal: model events' temporal property distributions
  - Duration, Frequency, Typical Time



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

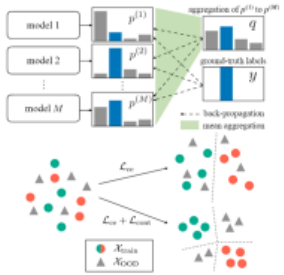
29

Indirect & weak supervision via other tasks, other resources

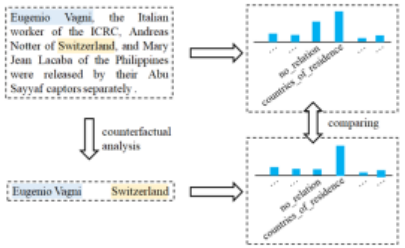
Robust Learning and Inference for IE

Agenda

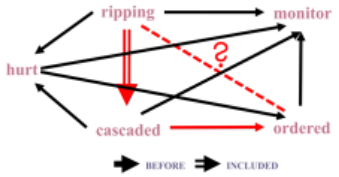
1. Noise-robust IE



2. Faithful IE



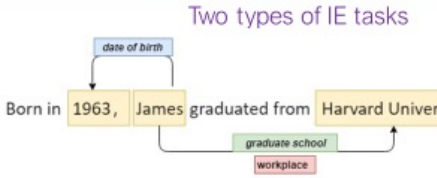
3. Logically Consistent IE



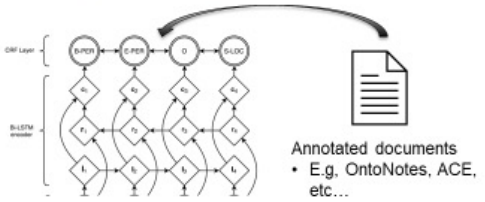
4. Open Research Directions

Set an agenda for robust, faithful, and consistent integrated IE

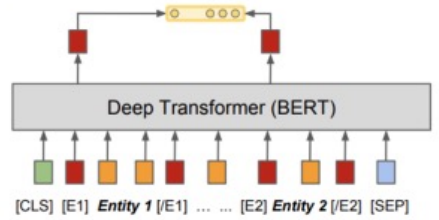
Two Types of IE Tasks



Lexical IE: Identifying **concepts** (entities, events, terms, etc.) and their **types**.



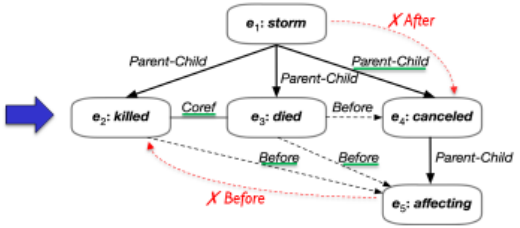
Relational IE: Identifying **relations** (or properties) of concepts



Consistency of IE

How do we ensure the extracts are **globally consistent**?

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.



Take event-event relation extraction as an example
- Temporal Relations
- Subevent Relations (Memberships)
- Event Coreference

Extracts are not independent, but a structure with dependencies
- E.g., Temporal relations cannot be a loop
- A main event cannot happen after a subevent

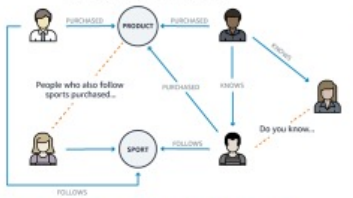
Li et al.A Logic-Driven Framework for Consistency of Neural Models. **EMNLP 2019**
Wang et al. Joint Constrained Learning for Event-Event Relation Extraction. **EMNLP 2020**

AI Needs to Understand Relations of Concepts

QA & Semantic Search

Relations of Entities

E-Commerce



Properties and Relations of Products

Comp. Bio. Med.

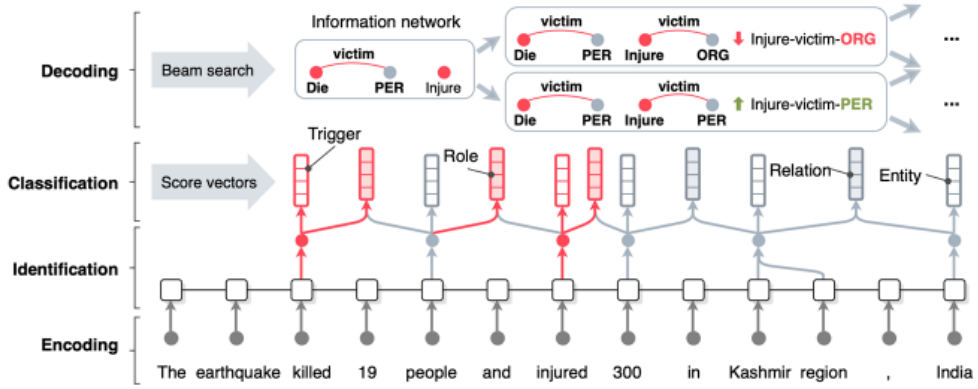
Interactions of (bio)molecules
Relations of diseases and drugs

IE automatically extracts structural knowledge about concepts and relations

Knowledge-guided IE

OneIE: Justify whether the entire graph makes sense

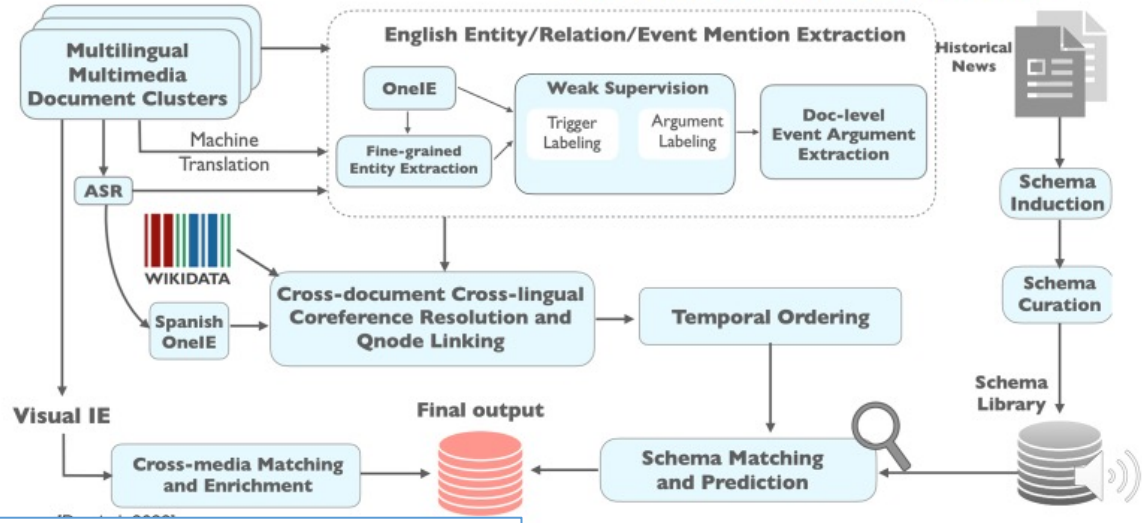
- OneIE framework extracts the information graph (nodes: entities and events, edges: relations and arguments) from a given sentence. (Lin et al., 2020)
- Main challenge for Joint IE: How to capture interactions between knowledge elements?



[Lin et al, 2020]

Schema-Guided Event Prediction: RESIN-11

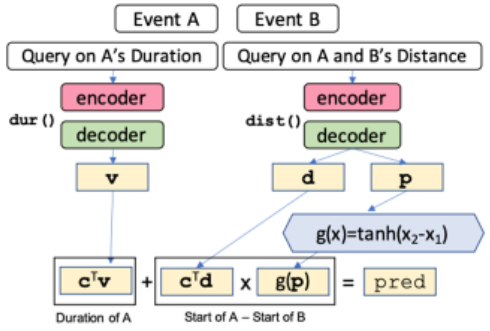
Dockerized system publicly available at Github: https://github.com/RESIN-KAIROS/RESIN-11



How various knowledge sources can be used to guide consistent IE

Commonsense Knowledge Enhanced IE

- Key idea: Leveraging duration



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

System	Start	End	All	Story
Majority	57.3	69.8	64.1	18.1
BiLSTM	53.7	63.5	59.1	10.9
Roberta-Large	78.5	78.3	78.4	26.1
T5-3B	79.4	77.4	78.3	26.9
BaseLM (T5-large)	75.5	75.4	75.4	22.6
BaseLM-MATRES	76.7	76.3	76.5	25.3
PtnTime (ours)	81.4	77.5	79.3	31.0
SymTime (ours)	**82.1**	**79.4**	**80.6**	**32.0**
SymTime-ZeroShot	77.0	73.1	74.9	21.6

Transferability of IE Systems

Why Transferability is Important

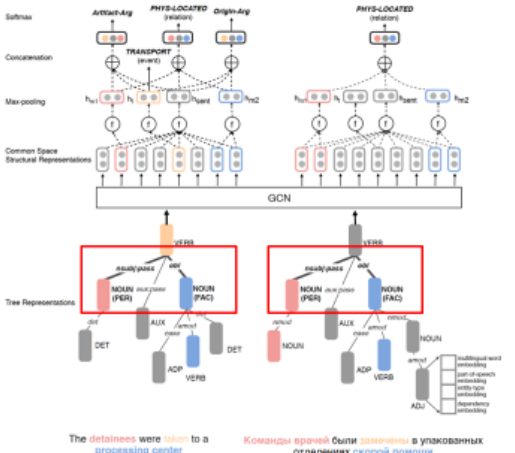
- Current status of information extraction
  - Domains: news, biomedical, clinical, legal, agriculture
  - Languages: English, Chinese, Spanish, Arabic
  - Number of Target Types: 3-100+ for entity recognition, ~100 for relation extraction, 33/38 for event extraction
- However, for other languages and domains, learning resources are insufficient.

High-resource domains — Low-resource domains

Cross-lingual Transfer: Language-agnostic Feature Representations

- Leveraging language-universal structural feature representations, e.g., dependency structures

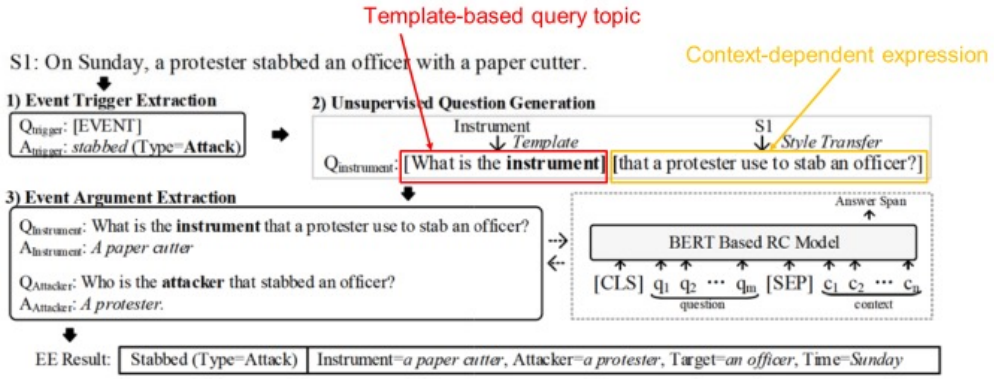
Transfer across tasks, languages, and continual learning for IE



connected in the dependency tree

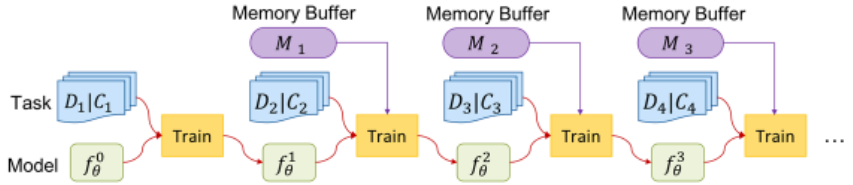
Cross-type Transfer: QA-based Event Extraction

- Questions can also be automatically generated in an unsupervised way (Liu et al., 2020)



Continual Learning for IE

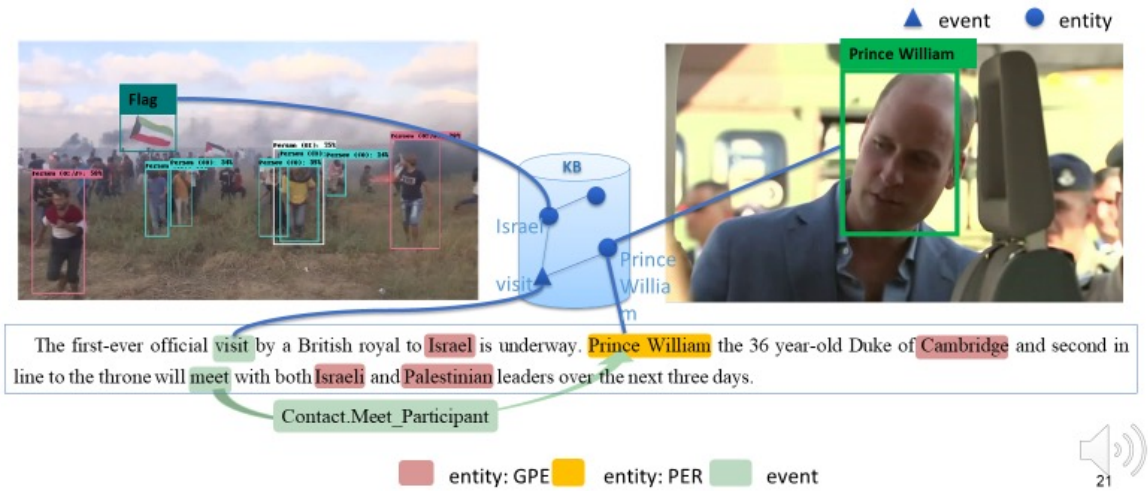
- How to mitigate the catastrophic forgetting?
  - Experience Replay: store K exemplars from old tasks into a memory and replay them periodically to prevent model forgetting previous knowledge when it's being trained on a new task
  - Knowledge Distillation: if a model extracts similar features or makes similar predictions for the same input as the old model, we can assume it preserves the knowledge
  - Task-specific Adapter: incrementally adding task-specific tunable parameters for each new task while fixing other parameters



Cross-Modal IE

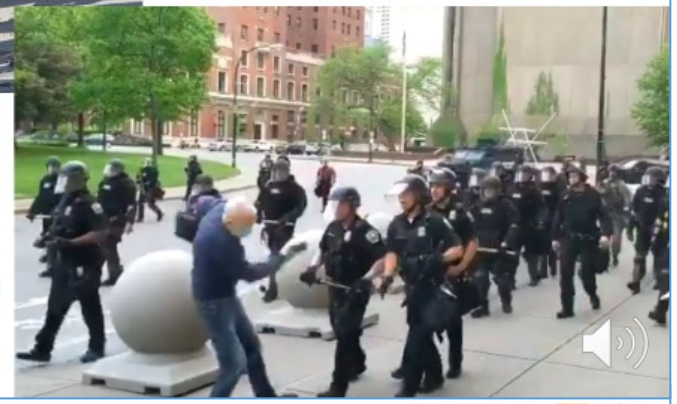
Multimedia Information Extraction

- Multimedia Knowledge Base with entities, relations and events.



Argument

Event	Attacking
Attacker	protesters
Target	police



Event	Attacking
Attacker	police
Target	protester

Event-level Cross-media Alignment

Investigators **inspect** parts of a destroyed car at the site of a car bombing in Beirut, Jan. 21, 2014.

Multimedia IE – challenges and opportunities

Visual Entity Linking and Coreference

- Flag recognition

US Flag

Missed detection

Different shape and angle
Partial observe

Ukraine Flag

Low resolution

Russia Flag (X) US Flag UK Flag

Ethics and Privacy

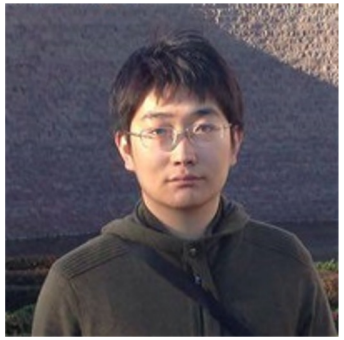
- Most of the technological advances carry risks with them.
- NLP and IE are no different, and perhaps even more sensitive than other technologies in today's information polluted world, and when privacy is left to individuals to deal with.
- In the absence of such regulation, society relies on those who apply technology to ensure that they apply it safely and in a transparent way.

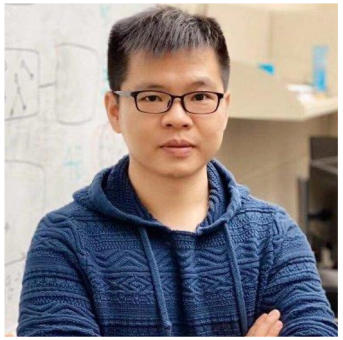
Tutorial Outline

- Introduction — 20 min.
  - Dan Roth
- Minimally and Indirectly Supervised IE — 35 min.
  - Ben Zhou
- Robust Learning and Inference for IE — 35 min.
  - Muhao Chen
- Break — 30 min.

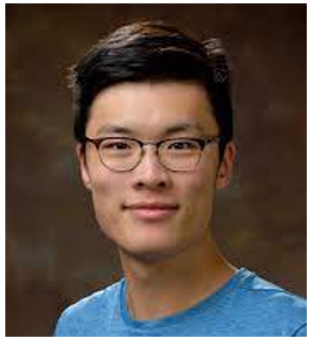
- Knowledge-guided IE — 15 min.
  - Manling Li
- Transferability of IE Systems — 35 min.
  - Lifu Huang
- Cross-Modal IE — 20 min.
  - Manling Li
- Conclusion and Future Work — 30 min.
  - Heng Ji, Dan Roth

New Frontiers of Information Extraction


Muhao Chen


Lifu Huang


Manling Li


Ben Zhou


Heng Ji


Dan Roth

July 2022

NAACL Tutorial

New Frontiers of Information Extraction

NAACL 2022

Challenges (1)

- Analyze Suspicious Transaction Reports (STRs) that banks submit to government agencies.
  - Extract relationships and transaction details.

- ALERTED ACTIVITY
  - *On 01/01/2011 the 01/01/2011 wire was stopped by ABC DE's filter:*
  - *JJ sent a $4,950.00 wire from account 123567 at GH Bank DE, through ABC DE, to account 123456 at Peoples' Bank of Malibu head office, Malibu, to be remitted to People's Bank of Malibu, XYZ Branch, for the benefit of BB for "Prayer Religious Items."*

- Determine: who did what to whom, where and when?
  - What are the transactions; from whom, to whom?
  - People involved, roles and affiliations
  - Organizations and their roles
  - Timelines of events

Challenges (2)

- Inter-banking rate issues
  - A client investigates if bankers were fixing the inter-banking rate.

  - *"Mate, can you raise the main one by 0.2 till Wednesday? I owe you a drink."*

- Messages here are typically very short; the language is colloquial and ungrammatical. Messages include many quantities (rates), how much to adjust, etc.

- A "simple" classification problem – but what is the label?
  - And, where is the training data?