Information Extraction

Dr. Muneendra Ojha

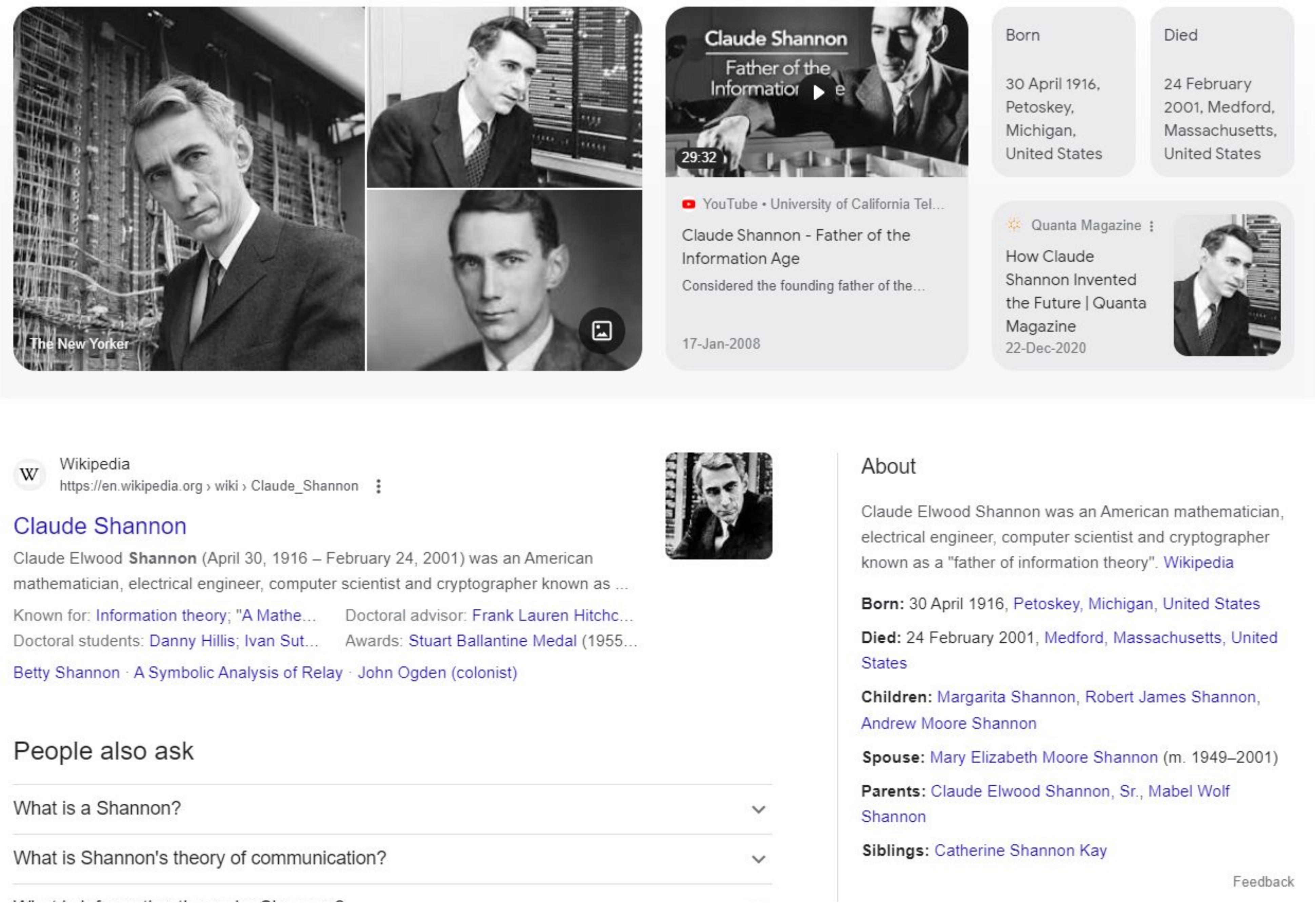
Assistant Professor

Department of Information Technology

Indian Institute of Information Technology - Allahabad

Information extraction

- Information extraction (IE) refers to the NLP task of extracting relevant information from text documents.
    - Example: the short blurbs we see when we search for a popular figure’s name on Google.



YouTube • University of California Tel...

Claude Shannon - Father of the Information Age

Considered the founding father of the...

17-Jan-2008

Born

30 April 1916, Petoskey, Michigan, United States

Died

24 February 2001, Medford, Massachusetts, United States

Quanta Magazine

How Claude Shannon Invented the Future | Quanta Magazine

22-Dec-2020

Wikipedia
https://en.wikipedia.org › wiki › Claude_Shannon

Claude Shannon

Claude Elwood **Shannon** (April 30, 1916 – February 24, 2001) was an American mathematician, electrical engineer, computer scientist and cryptographer known as ...

Known for: Information theory; "A Mathe... Doctoral advisor: Frank Lauren Hitchc...
Doctoral students: Danny Hillis; Ivan Sut... Awards: Stuart Ballantine Medal (1955...

Betty Shannon · A Symbolic Analysis of Relay · John Ogden (colonist)

People also ask

What is a Shannon?

What is Shannon's theory of communication?

About

Claude Elwood Shannon was an American mathematician, electrical engineer, computer scientist and cryptographer known as a "father of information theory". Wikipedia

Born: 30 April 1916, Petoskey, Michigan, United States

Died: 24 February 2001, Medford, Massachusetts, United States

Children: Margarita Shannon, Robert James Shannon, Andrew Moore Shannon

Spouse: Mary Elizabeth Moore Shannon (m. 1949–2001)

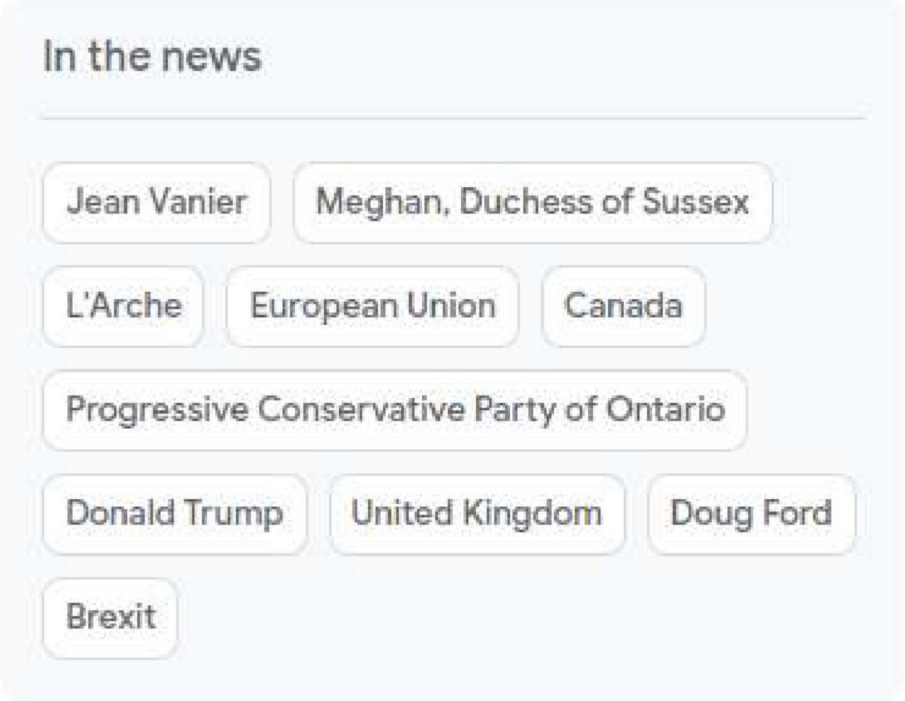
Parents: Claude Elwood Shannon, Sr., Mabel Wolf Shannon

Siblings: Catherine Shannon Kay

Feedback

IE Applications

- Tagging news and other content
  - It's useful if some texts (e.g. news snippets) are tagged with important entities mentioned within them.



  - People (e.g., Jean Vanier), organizations (e.g., Progressive Conservative Party of Ontario), locations (e.g., Canada), and events (e.g., Brexit) currently in the news are extracted and shown to the reader so that they can go directly to news about a specific entity.

IE Applications

- Chatbots
  - A chatbot needs to understand the user’s question in order to generate/retrieve a correct response.
  - For example, consider the question, “What are the best cafes around the Eiffel Tower?”
  - The chatbot needs to understand that “Eiffel Tower” and “cafe” are locations, then identify cafes within a certain distance of the Eiffel Tower.

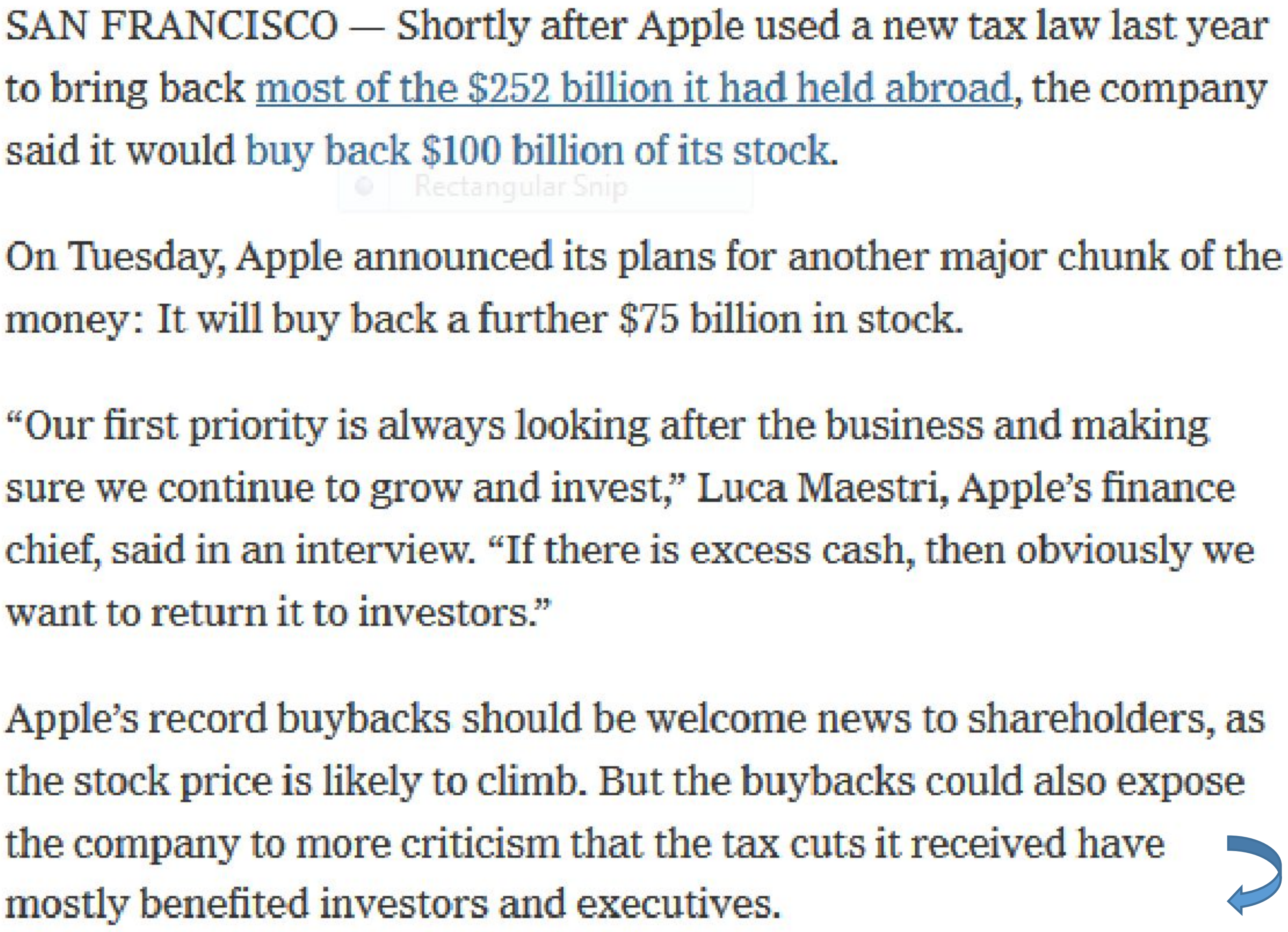
IE Applications

- Applications in social media
  - A lot of information is disseminated through social media channels like Twitter
  - Extracting informative excerpts from social media text may help in decision making.

IE Tasks

- The overarching goal of IE is to extract “knowledge” from text, and each of IE tasks provides different information.

What information you can draw?



SAN FRANCISCO — Shortly after Apple used a new tax law last year to bring back most of the $252 billion it had held abroad, the company said it would buy back $100 billion of its stock.

Rectangular Snip

On Tuesday, Apple announced its plans for another major chunk of the money: It will buy back a further $75 billion in stock.

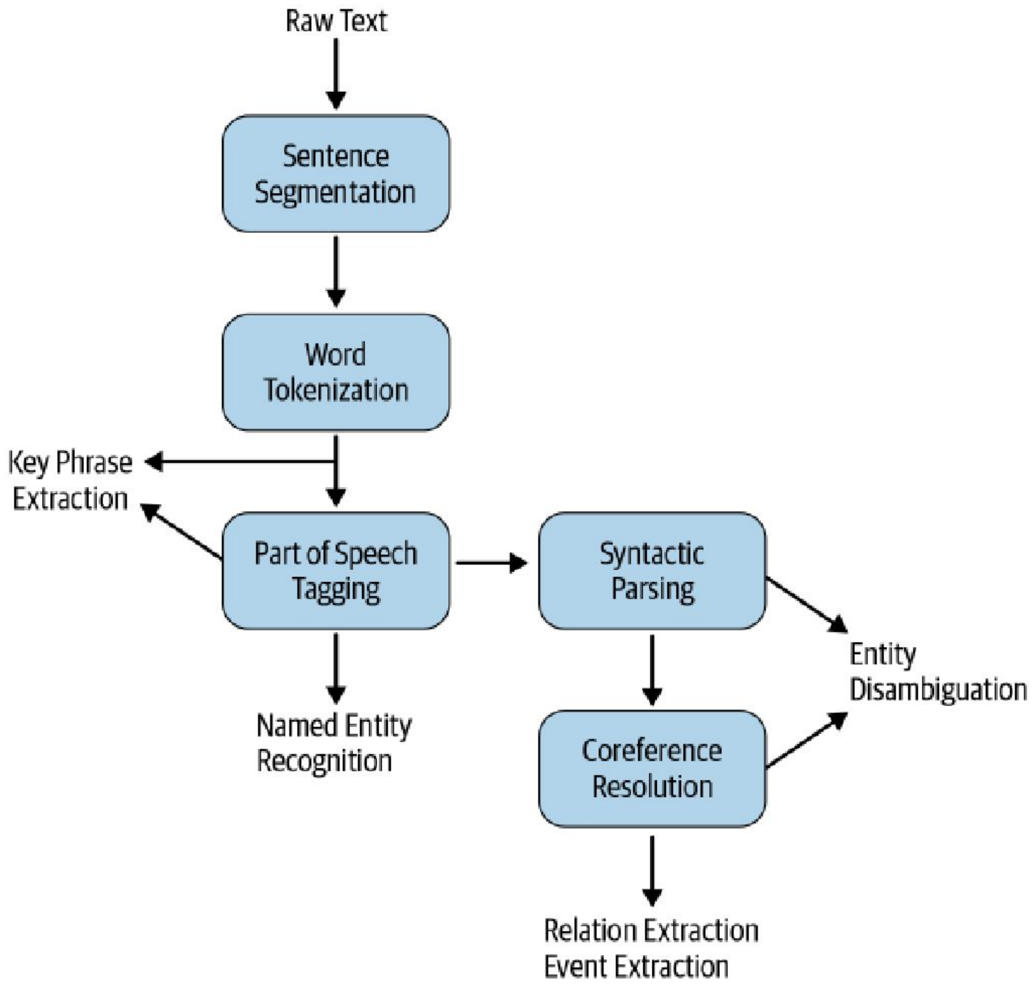
“Our first priority is always looking after the business and making sure we continue to grow and invest,” Luca Maestri, Apple’s finance chief, said in an interview. “If there is excess cash, then obviously we want to return it to investors.”

Apple’s record buybacks should be welcome news to shareholders, as the stock price is likely to climb. But the buybacks could also expose the company to more criticism that the tax cuts it received have mostly benefited investors and executives.

IE Tasks

- Identifying that the article is about “buyback” or “stock price” relates to the task of ***keyword*** or ***keyphrase extraction (KPE)***.
- Identifying Apple as an organization and Luca Maestri as a person comes under the task of ***named entity recognition (NER)****.*
- Recognizing that Apple is not a fruit, but a company and that it refers to Apple, Inc. and not some other company with the word “apple” in its name is the task of ***named entity disambiguation and linking (NED/NEL).***
- Extracting the information that Luca Maestri is the finance chief of Apple refers to the IE task of ***relation extraction***.

The General Pipeline for IE



Keyphrase Extraction (KPE)

- Keyword and phrase extraction is the IE task of extracting important words and phrases that capture the gist of the text from a given text document.
- It's useful for several NLP tasks, such as search/information retrieval, automatic document tagging, recommendation systems, text summarization, etc.

Keyphrase Extraction (KPE)

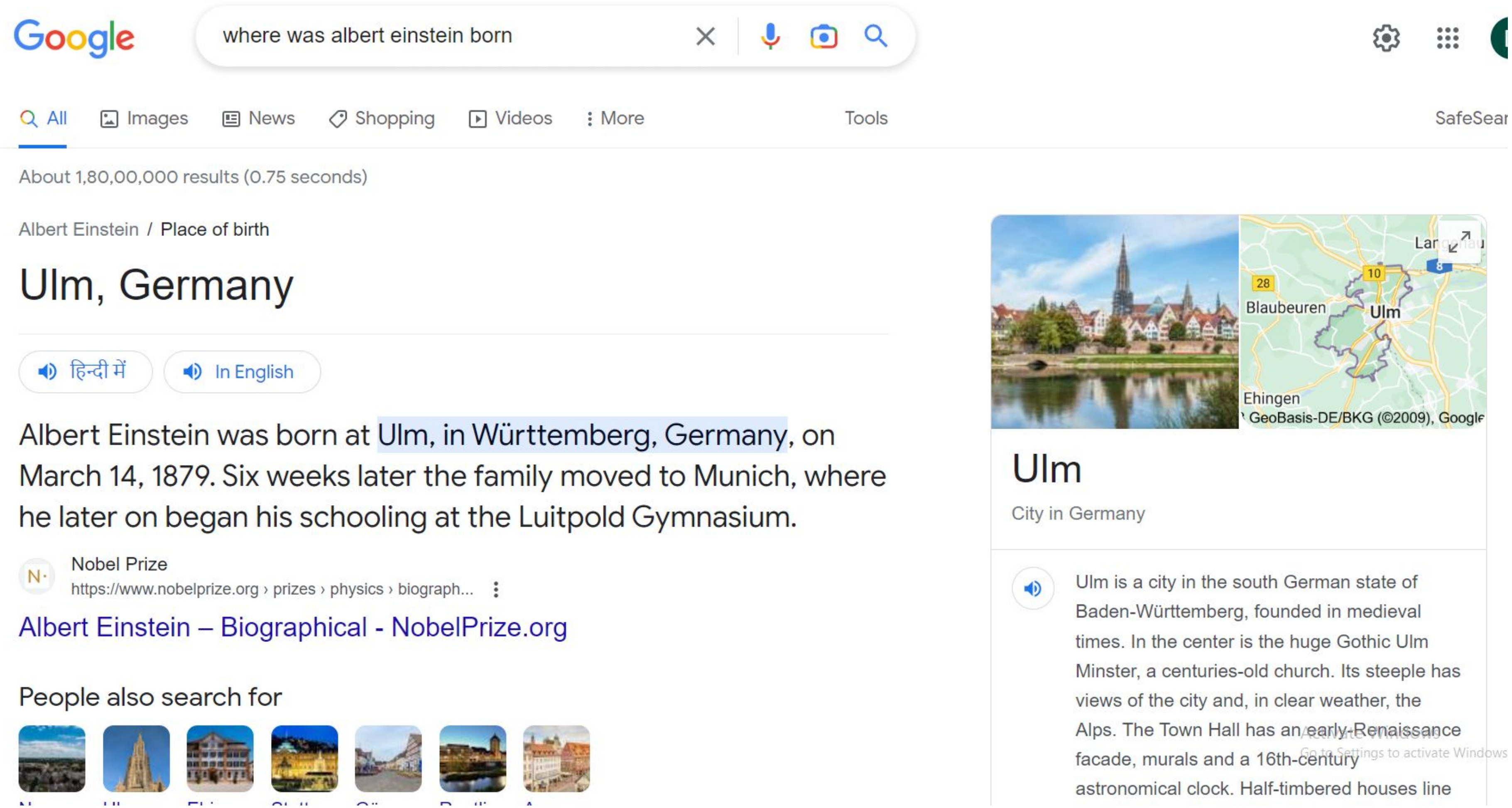
- Two most commonly used methods for KPE to are ***supervised learning*** and ***unsupervised learning***.
- Supervised learning approaches require corpora with texts and their respective keyphrases and use engineered features or DL techniques.
  - Creating such labeled datasets for KPE is a time- and cost-intensive endeavor.
- Unsupervised approaches that do not require a labeled dataset and are largely domain agnostic are more popular for KPE.
  - These approaches are also more commonly used in real-world KPE applications.
- Recent research has shown that state-of-the-art DL methods for KPE don't perform any better than unsupervised approaches.

Keyphrase Extraction (KPE)

- Majority of unsupervised KPE algorithms represent the words and phrases in a text as nodes in a weighted graph
    - The weight indicates the importance of that keyphrase.
    - Keyphrases are then identified based on how connected they are with the rest of the graph.
    - The top-N important nodes from the graph are then returned as keyphrases.
- Important nodes are those words and phrases that are frequent enough and also well connected to different parts of the text.

Named Entity Recognition (NER)

- “Where was Albert Einstein born?”—using Google search



About 1,80,00,000 results (0.75 seconds)

Albert Einstein / Place of birth

Ulm, Germany

Albert Einstein was born at Ulm, in Württemberg, Germany, on March 14, 1879. Six weeks later the family moved to Munich, where he later on began his schooling at the Luitpold Gymnasium.

Nobel Prize
https://www.nobelprize.org › prizes › physics › biograph...

Albert Einstein – Biographical - NobelPrize.org

People also search for

Ulm

City in Germany

Ulm is a city in the south German state of Baden-Württemberg, founded in medieval times. In the center is the huge Gothic Ulm Minster, a centuries-old church. Its steeple has views of the city and, in clear weather, the Alps. The Town Hall has an early-Renaissance facade, murals and a 16th-century astronomical clock. Half-timbered houses line

Named Entity Recognition (NER)

- NER refers to the IE task of identifying the entities in a document.
- Entities are typically names of persons, locations, and organizations, and other specialized strings, such as money expressions, dates, products, names/numbers of laws or articles, and so on.

Building an NER System

- One approach is to maintain a large collection of person/ organization/location names that are the most relevant to our domain;
  - Such a collection is called a gazetteer.
- ***Problem:*** How to deal with new names? How to update the database? How to keep track of aliases (e.g., USA, United States, etc.)?

Building an NER System

- Another approach is based on a compiled list of patterns based on word tokens and POS tags.
- For example, a pattern “NNP was born,” where “NNP” is the POS tag for a proper noun, indicates that the word that was tagged “NNP” refers to a person.
- Such rules can be programmed to cover as many cases as possible to build a rule-based NER system.
- Stanford NLP’s RegexNER [19] and spaCy’s EntityRuler [20] provide functionalities to implement rule-based NER.

Building an NER System

- A more practical approach to NER is to train an ML model, which can predict the named entities in unseen text.
- For each word, a decision has to be made whether or not that word is an entity, and if it is, what type of entity it is.
- This approach is very similar to classification problems.
- NER is traditionally modeled as a ***sequence classification problem***, where the entity prediction for the current word also depends on the context.

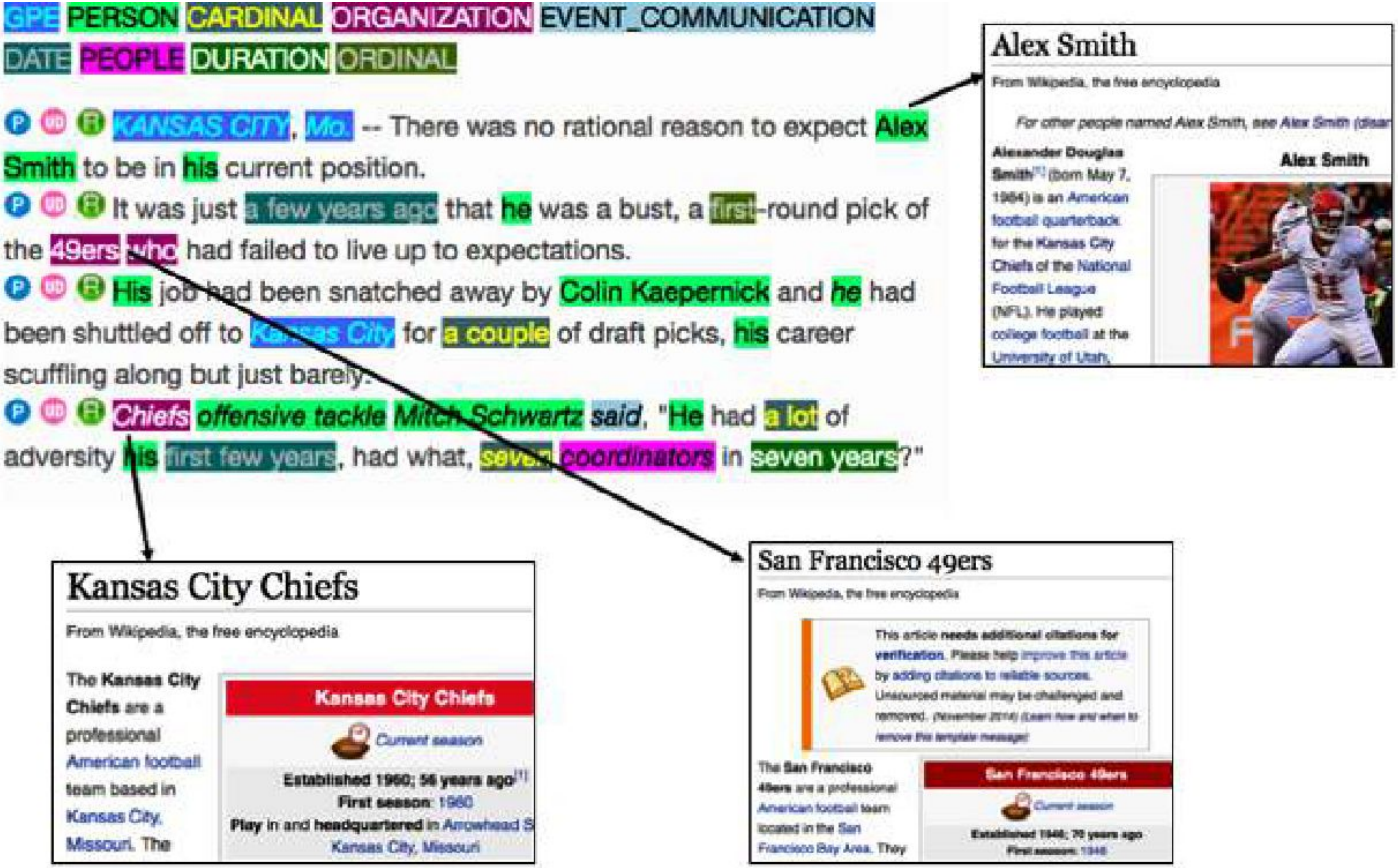
Normal classifier vs Sequence classifier

Washington is a rainy state

- A normal classifier has to make a decision as to whether Washington refers to a person or the State of Washington without looking at the surrounding words.
- However, to classify the word “Washington” in this sentence as a location one needs to look at the context in which it’s being used.
- Thus sequence classifiers are used for training NER models.

Named Entity Disambiguation and Linking

- Consider a scenario for a large newspaper publication.
  - We’re asked to build a system that creates a visual representation of news stories by connecting different entities mentioned in the stories to what they refer to in the real world.
- Doing this requires knowledge of several IE tasks beyond NER and KPE.
- As a first step, we have to know what these entities or keywords actually refer to in the real world.



GPE PERSON CARDINAL ORGANIZATION EVENT_COMMUNICATION
DATE PEOPLE DURATION ORDINAL

P UD R KANSAS CITY, Mo. -- There was no rational reason to expect Alex Smith to be in his current position.

P UD R It was just a few years ago that he was a bust, a first-round pick of the 49ers who had failed to live up to expectations.

P UD R His job had been snatched away by Colin Kaepernick and he had been shuttled off to Kansas City for a couple of draft picks, his career scuffling along but just barely.

P UD R *Chiefs offensive tackle Mitch Schwartz said*, "He had a lot of adversity his first few years, had what, seven coordinators in seven years?"

Alex Smith

From Wikipedia, the free encyclopedia

For other people named Alex Smith, see Alex Smith (disar

Alexander Douglas Smith[1] (born May 7, 1984) is an American football quarterback for the Kansas City Chiefs of the National Football League (NFL). He played college football at the University of Utah,

Alex Smith

Kansas City Chiefs

From Wikipedia, the free encyclopedia

The **Kansas City Chiefs** are a professional American football team based in Kansas City, Missouri. The

Kansas City Chiefs

Current season

Established 1960; 56 years ago[1]
First season: 1960
Play in and **headquartered** in Arrowhead S
Kansas City, Missouri

San Francisco 49ers

From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(November 2014) (Learn how and when to remove this template message)*

The **San Francisco 49ers** are a professional American football team located in the San Francisco Bay Area. They

San Francisco 49ers

Current season

Established 1946; 70 years ago
First season: 1946

Named Entity Disambiguation and Linking

"Lincoln drives a Lincoln Aviator and lives on Lincoln Way."

- All three mentions of "Lincoln" refer to different entities and different types of entities:
  - the first is a person
  - the second one is a vehicle
  - the third is a location
- How can we reliably link the three Lincolns to their correct Wikipedia pages? All three mentions of "Lincoln" here refer to different entities and different
- types of entities: the first Lincoln is a person, the second one is a vehicle, and
- the third is a location. How can we reliably link the three Lincolns to their correct
- Wikipedia pages

Named Entity Disambiguation and Linking

- Named entity disambiguation (NED) refers to the NLP task of:

  ***Assigning a unique identity to entities mentioned in the text*.**

- It’s also the first step in moving toward more sophisticated tasks to address the scenario mentioned above by identifying relationships between entities.

Named Entity Disambiguation and Linking

- NER and NED together are known as named entity linking (NEL).
- Some other NLP applications that would need NEL include question answering and constructing large knowledge bases of connected events and entities, such as the Google Knowledge Graph
- ***Problem:*** How do we build an IE system for performing NEL?

Named Entity Disambiguation and Linking

- As NER identifies entities and their spans using contextual information encoded by a range of features, NEL also relies on context.
- It requires going beyond POS tagging in terms of the NLP pre-processing needed.
- At a minimum, NEL needs some form of parsing to identify linguistic items like subject, verb, and object.
- It may also need coreference resolution to resolve and link multiple references to the same entity (e.g., Albert Einstein, the scientist, Einstein, etc.) to the same reference in a large, encyclopedic knowledge base (e.g., Wikipedia).
- This is typically modeled as a supervised ML problem and evaluated in terms of precision, recall, and F1 scores on standard test sets.

Named Entity Disambiguation and Linking

- State-of-the-art NEL uses a range of different neural network architectures.
- Clearly, learning an NEL model requires the presence of a large, annotated dataset as well as some kind of encyclopedic resource to link to.
- Further, NEL is a much more specialized NLP task compared to what we've seen so far (text representation, text classification, NER, KPE).
- Industry practitioners use off-the-shelf, pay-as-you-use services offered by big providers such as IBM (Watson) and Microsoft (Azure) for NEL rather than developing an in-house system.

Relationship Extraction (RE)

- Suppose a company mines multiple news articles to derive financial insights.
- To do such analysis on thousands of news texts every day, it needs a constantly updated knowledge base connecting different people, organizations, and events based on the news content.
- ***Problem:*** How will we get started building such a tool?

Relationship Extraction (RE)

- The IE tasks seen so far—KPE, NER, and NEL—are all useful to a certain extent in helping identify entities, events, keyphrases, etc.
- But how do we go from there to the next step of “connecting” them with some relation? What exactly are the relations? How will we extract them?

Relationship Extraction (RE)

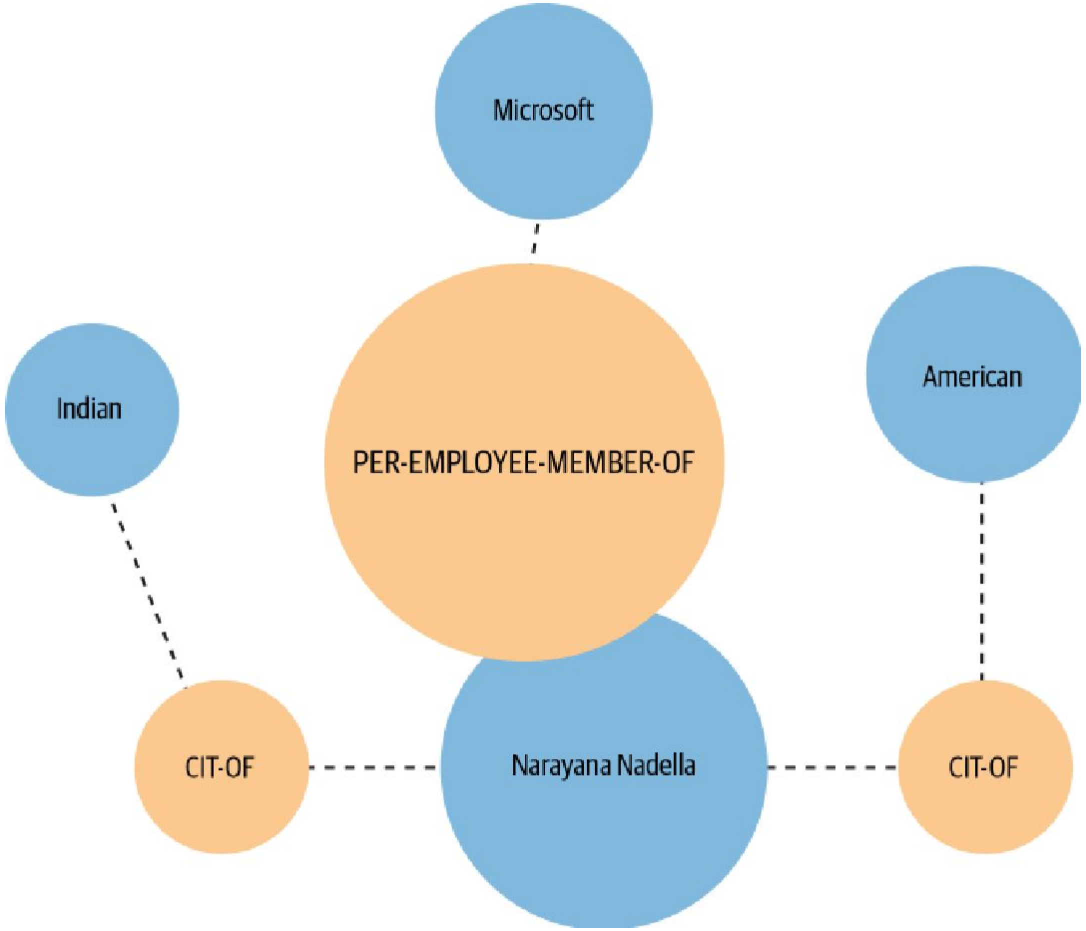
- Let’s revisit the news article showing a screenshot of a New York Times article.
- One relation that can be extracted is: (Luca Maestri, finance chief, Apple).
- Here, we connect Luca Maestri to Apple with the relationship of finance chief.

Relationship Extraction (RE)

- Relationship extraction (RE) is the IE task that deals with extracting entities and relationships between them from text documents.
- It’s an important step in building a knowledge base, and it’s also useful in improving search and developing question-answering systems.

Relationship extraction (RE)

Satya Narayana Nadella is an Indian-American business executive. He currently serves as the Chief Executive Officer (CEO) of Microsoft, succeeding Steve Ballmer in 2014. Before becoming chief executive, he was Executive Vice President of Microsoft's Cloud and Enterprise Group, responsible for building and running the company's computing platforms.



Relationship extraction (RE)

- The output shows that Narayana Nadella is a person related to Microsoft as an employee, related to India and America as a citizen, and so on.
- ***Problem:*** How does one proceed with extracting such relationships from a piece of text?

Relationship extraction (RE)

- Apart from identifying entities and disambiguating them, one needs to model the process of extracting the relationships between them by considering the words connecting the entities in a sentence, their sense of usage, and so on.
- ***Problem:*** What constitutes a “relation”?

Relationship extraction (RE)

- Relations can be specific to a given domain
  - Medical domain, relations could include type of injury, location of injury, cause of injury, treatment of injury, etc.
  - Financial domain, relations could mean something completely different.
  - A few generic relations between people, locations, and organizations are: located in, is a part of, founder of, parent of, etc.
- How do we extract them?

Approaches to RE

- In NLP, RE is a well-researched topic, and—starting from handwritten patterns to different forms of supervised, semi-supervised, and unsupervised learning—various methods have been explored (and are still being used) for building RE systems.

Approaches to RE

- Hand-built patterns consist of regular expressions that aim to capture specific relationships.
- For example, a pattern such as “PER, [something] of ORG” can indicate a sort of “is-a-part-of ” relation between that person and organization.
- Such patterns have the advantage of high precision, but they often have less coverage, and it could be challenging to create such patterns to cover all possible relations within a domain.

Approaches to RE

- The datasets used to train RE systems contain a set of predefined relations, similar to classification datasets.
- This consists of modeling it as a two-step classification problem:
  1. Whether two entities in a text are related (binary classification).
  2. If they are related, what is the relation between them (multiclass classification)?

Approaches to RE

- These are treated as a regular text classification problem, using handcrafted features, contextual features like in NER (e.g., words around a given entity), syntactic structure (e.g., a pattern such as NP VP NP, where NP is a noun phrase and VP is a verb phrase), and so on.
- Neural models typically use different embedding representations followed by an architecture like recurrent neural networks

Approaches to RE

- Both supervised approaches and pattern-based approaches are domain-specific.
- Getting large amounts of annotated data each time we start on a new domain is challenging and expensive.
- Bootstrapping can be used in such scenarios, starting with a small set of seed patterns and generalizing by learning new patterns based on the sentences extracted using these seed patterns.

Approaches to RE

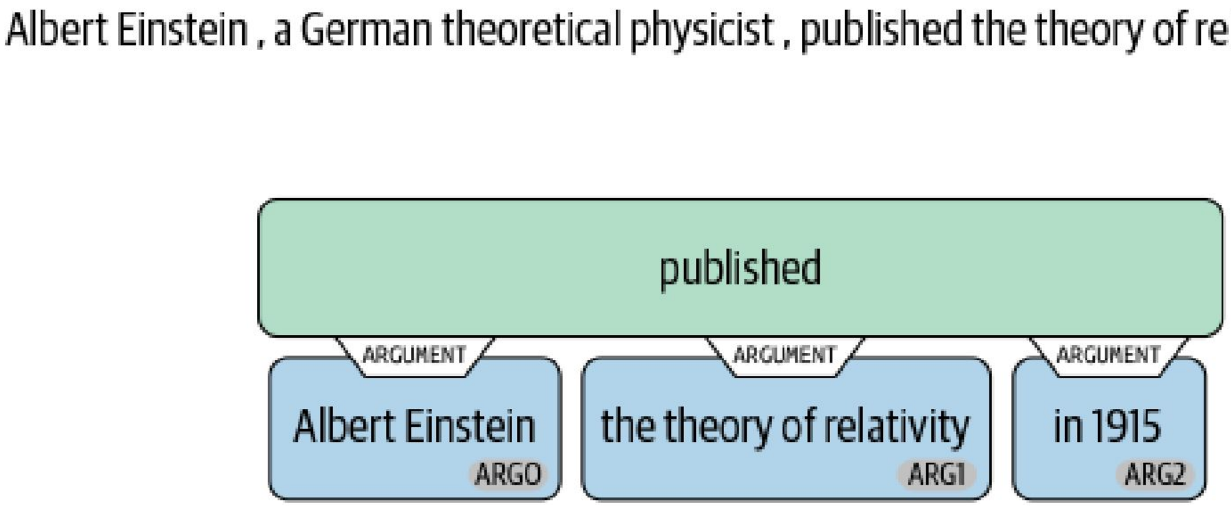
- In another approach, instead of using a small set of seed patterns, large databases such as Wikipedia, Freebase, etc., are used to first collect thousands of examples of many relations (e.g., using Wikipedia infoboxes), thereby creating a large dataset of relations.
- This can then be followed by a regular supervised relation extraction approach. Even this works only when such large databases exist.

Approaches to RE

- In scenarios where we cannot procure training data for supervised approaches, we apply unsupervised approaches.
- Unsupervised RE (also known as “open IE”) aims to extract relations from the web without relying on any training data or any list of relations.
- The relations extracted are in the form of <verb, argument1, argument2> tuples.
- Sometimes, a verb may have more arguments.

Example:

- We see the relation as a verb and its three arguments <published, albert einstein, the theory of relativity, in 1915>.
- We can also extract the relation tuples <published, albert einstein, the theory of relativity>, <published, albert einstein, in 1915>, and <published, theory of relativity, 1915>.

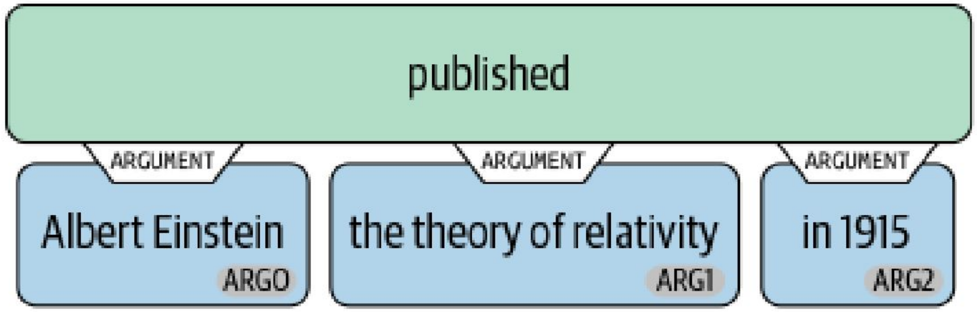


Albert Einstein , a German theoretical physicist , published the theory of relativity in 1915 .

Example:

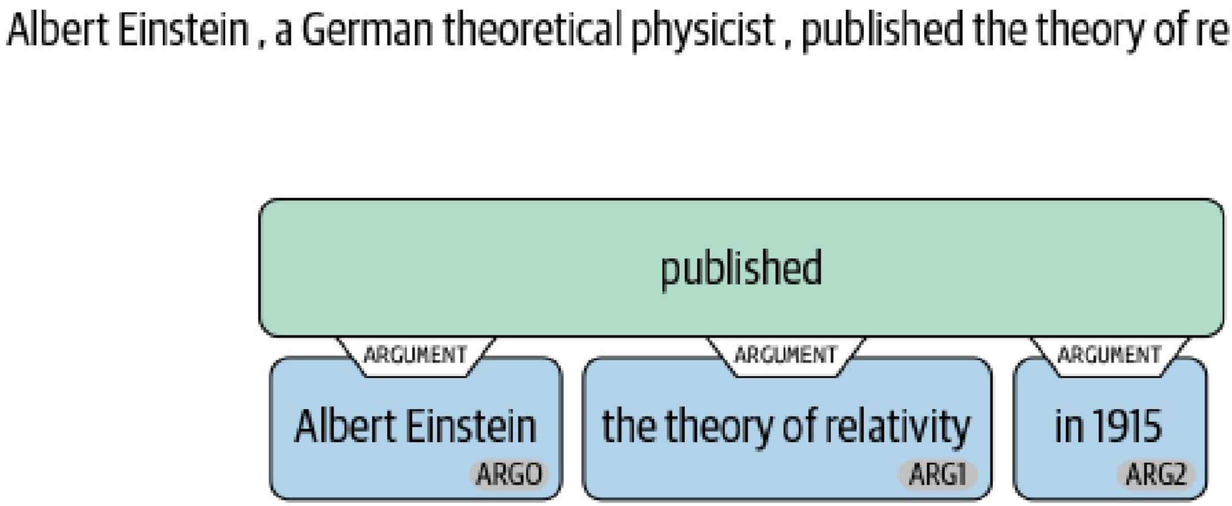
- In such a system, we see as many tuples/quadruples as the number of verbs.
- While this is an advantage, the challenge lies in mapping the extracted versions to some standardized set of relations (e.g., fatherOf, motherOf, inventorOf, etc.) from a database.

Albert Einstein , a German theoretical physicist , published the theory of relativity in 1915 .



Example:

- Then to extract specific relations from this information (if we need to), we would have to devise our own procedures combining the outputs of NER/NEL, coreference resolution, and open IE.



Albert Einstein , a German theoretical physicist , published the theory of relativity in 1915 .

Temporal Information Extraction

- ***Email text:*** "Let us meet at 3 p.m. today and decide on what to present at the meeting on Friday."
- ***Objective:*** To identify and populate calendars with events extracted from such conversations.
  - Apart from extracting date and time information (3 pm, today, Friday) from the text, we should also convert the extracted data into a standard form (e.g., mapping the expression "on Friday" to the exact date, based on context, and "today" to today's date).

Temporal Information Extraction

- While extracting date and time information can be done using a collection of handcrafted patterns in the form of regex, or by applying supervised sequence labeling techniques as we did for NER, normalization of extracted date and time into a standard date-time format can be challenging.
- Together, these tasks are referred to as temporal IE and normalization.
- Contemporary approaches to such temporal expression normalization are primarily rule-based and coupled with semantic analysis

Temporal Information Extraction

- Duckling is a Python library recently released by Facebook’s bots team that was used to build bots for Facebook Messenger.
- The package is designed to parse text and get structured data.
- Among the many tasks it can do, it can process the natural language text data to extract temporal events

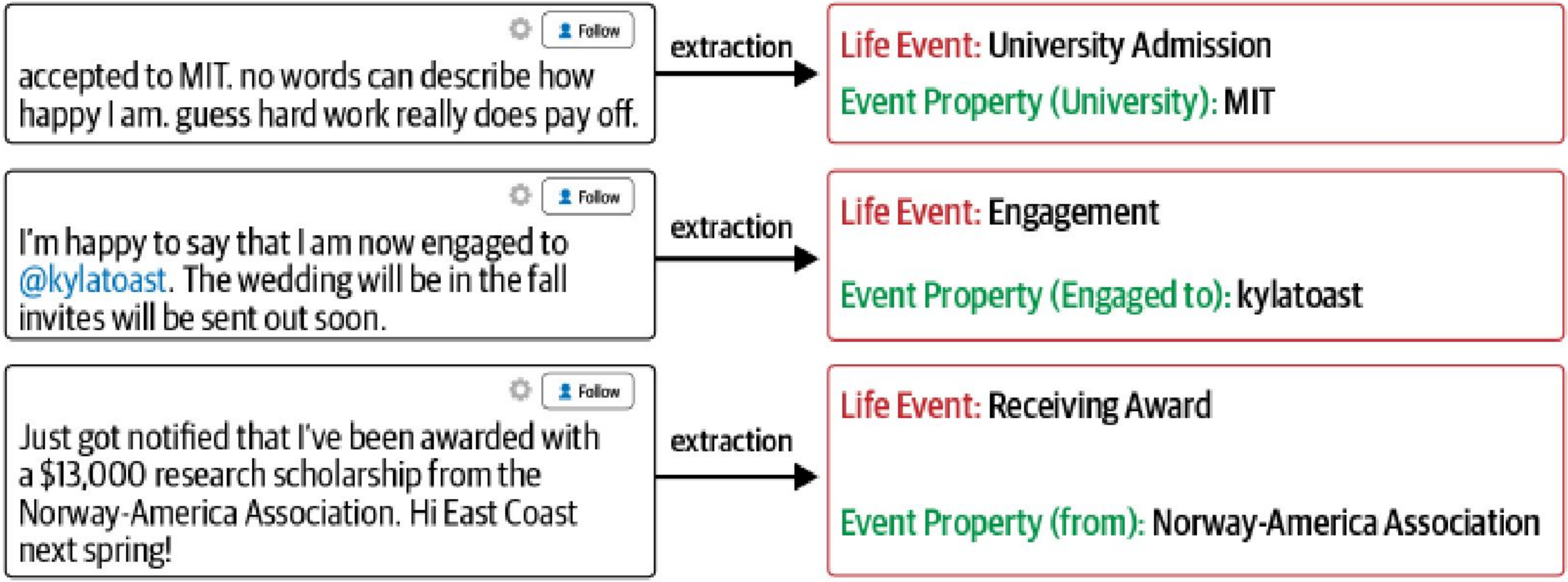
Event Extraction

- In the email-text example, the aim of extracting temporal expressions is to eventually extract information about an “event.”
- Events can be anything that happens at a certain point in time: meetings, increase in fuel prices in a region at a certain time, presidential elections, the rise and fall of stocks, life events like birth, marriage, and demise, and so on.
- Event extraction is the IE task that deals with identifying and extracting events from text data.

Event Extraction

- It’s believed that ***Bloomberg Terminal*** has a submodule that reports major events that are identified from thousands of news sources and social channels like Twitter in real-time across the globe.
- A popular, fun application of event extraction is the ***congratsbot***.
- The bot reads through tweets and responds with a “congrats” message if it sees any event that one should be congratulated on

Event Extraction



accepted to MIT. no words can describe how happy I am. guess hard work really does pay off.

Life Event: University Admission

Event Property (University): MIT

I'm happy to say that I am now engaged to @kylatoast. The wedding will be in the fall invites will be sent out soon.

Life Event: Engagement

Event Property (Engaged to): kylatoast

Just got notified that I've been awarded with a $13,000 research scholarship from the Norway-America Association. Hi East Coast next spring!

Life Event: Receiving Award

Event Property (from): Norway-America Association

Event Extraction

- Event extraction is treated as a supervised learning problem in NLP literature.
- Contemporary approaches use sequence tagging and multilevel classifiers, much like with relationship extraction.
- The ultimate goal is to identify various events over time periods, connect them, and create a temporally ordered event graph.

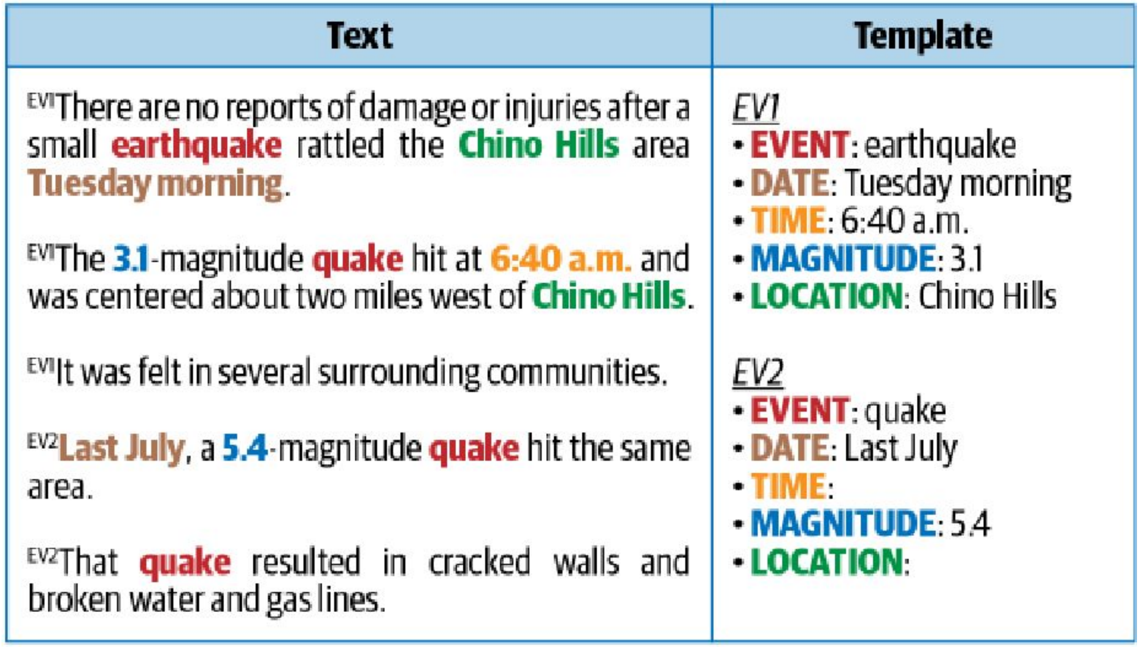
Event Extraction

- This is still an active area of research, and working solutions for event extraction like those mentioned previously only work for specific scenarios; i.e., there are no relatively generic solutions like we saw for RE, NER, etc.
- If you end up doing a project that requires event extraction, the best way forward is to first start with a rule-based approach based on domain knowledge, then follow it up with weak supervision. As you start accumulating more data, you can move toward ML approaches.

Template Filling

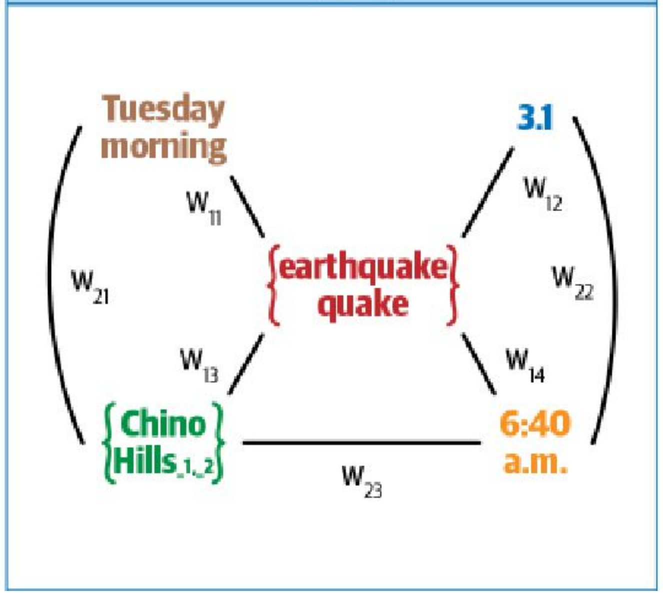
- In some application scenarios, such as weather forecasts and financial reports, the text format is fairly standard, and what changes are the specific details pertaining to that situation.
- Such scenarios are good use cases for an IE task called template filling, where the task is to model text generation as a slot-filling problem.

Template Filling



Text	Template
[EV1]There are no reports of damage or injuries after a small **earthquake** rattled the **Chino Hills** area **Tuesday morning**. [EV1]The **3.1**-magnitude **quake** hit at **6:40 a.m.** and was centered about two miles west of **Chino Hills**. [EV1]It was felt in several surrounding communities. [EV2]**Last July**, a **5.4**-magnitude **quake** hit the same area. [EV2]That **quake** resulted in cracked walls and broken water and gas lines.	*EV1* • **EVENT**: earthquake • **DATE**: Tuesday morning • **TIME**: 6:40 a.m. • **MAGNITUDE**: 3.1 • **LOCATION**: Chino Hills *EV2* • **EVENT**: quake • **DATE**: Last July • **TIME**: • **MAGNITUDE**: 5.4 • **LOCATION**:

Entity graph



Template Filling

- Generally, the templates to fill are pre-defined. This is typically modeled as a two stage, supervised ML problem, similar to relation extraction.
- The first step involves identifying whether a template is present in a given sentence, and the second step involves identifying slot fillers for that template, with a separate classifier trained for each slot.
- Work is being done in the direction of automatically inducing templates.

Assignment

- Chapter 5, Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems by Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, and Harshit Surana, O’Reilly Publications.
  - [55] Molumby, Conor and Joe Whitwell. “General Election 2019: Semi-Automation Makes It a Night of 689 Stories”. BBC News Labs, December 13, 2019.
  - [56] Reiter, Ehud. “Election Results: Lessons from a Real-World NLG System”, Ehud Reiter’s Blog, December 23, 2019.

Course Project – Choose one!

1. Build an Event Extractor
2. Build a Template Filler

References

- Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems by Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, and Harshit Surana, O’Reilly Publications.