Text Analytics & Business Application

Information Extraction

Qinglai He

Department of Operations and Information Management

Wisconsin School of Business

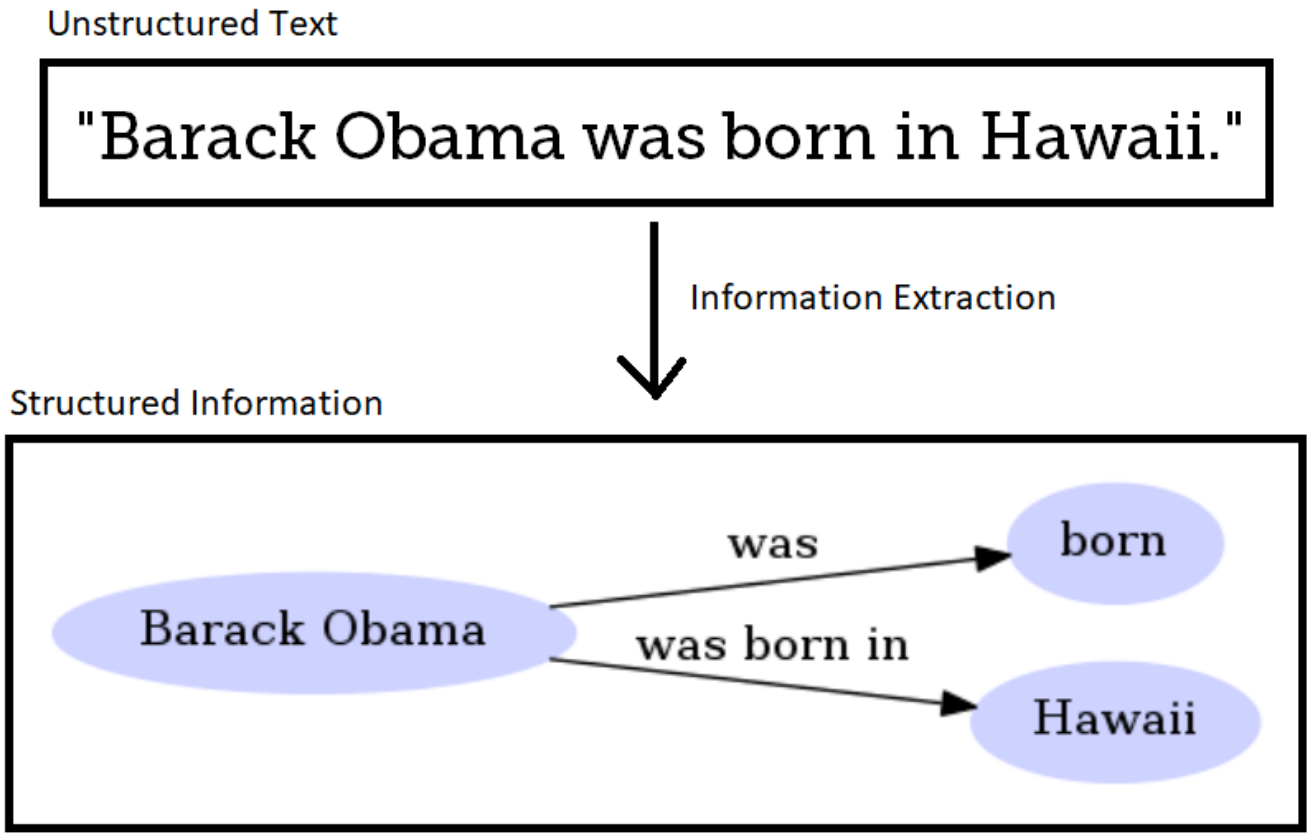
WISCONSIN SCHOOL OF BUSINESS

Outline of Today's Class

- IE applications
- IE tasks
- Key phrase extraction
- Name entity recognition and POS tagging
- Relationship extraction
- Other advanced IE tasks

IE Applications

How Does Information Extraction Work?

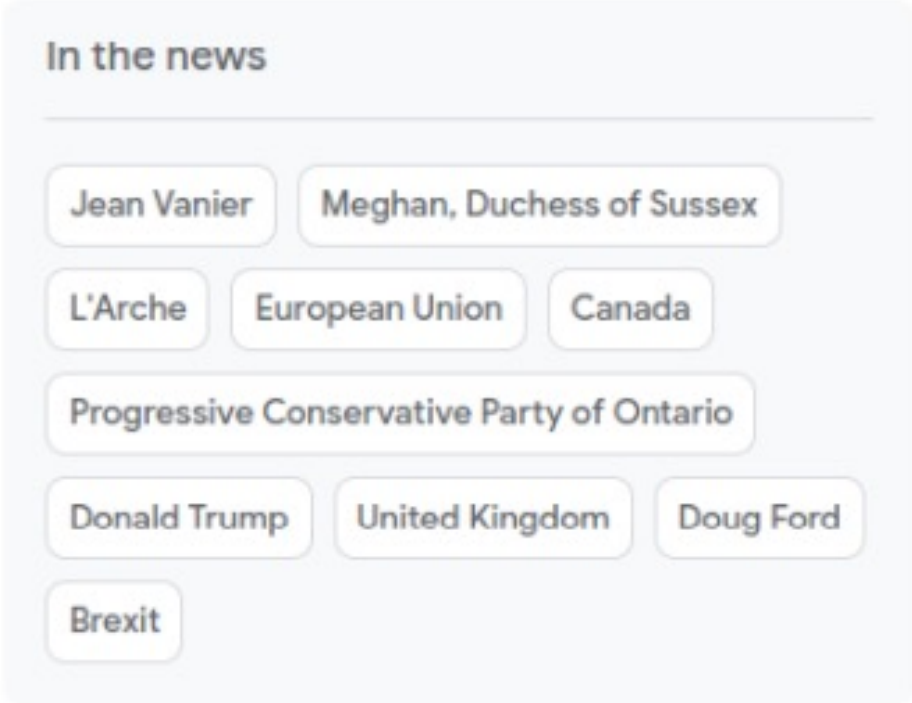


IE Applications

- IE is used in a wide range of real-world applications, from news articles, to social media, and even receipts.
- Here, we'll cover the details of a few of them:
  - Tagging news and other content
  - Chatbots
  - Applications in social media
  - Extracting data from forms and receipts

Tagging News and Other Content

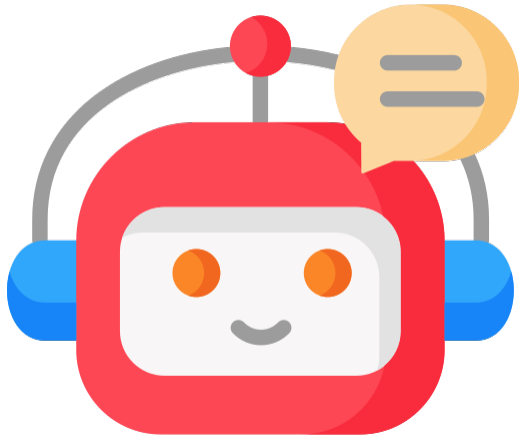
- Useful for applications such as search engines and recommendation systems.
- People, organizations, locations, and events currently in the news are extracted.
- Extracted entities are shown to the reader for direct access to news about a specific entity.

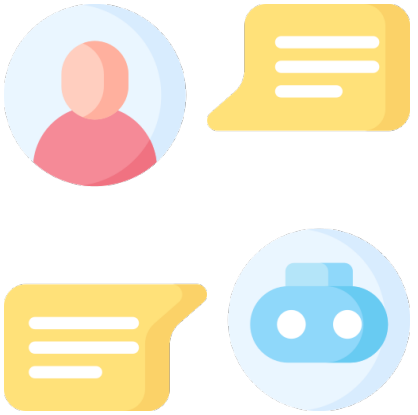


Screenshot from the Google News homepage

Chatbots

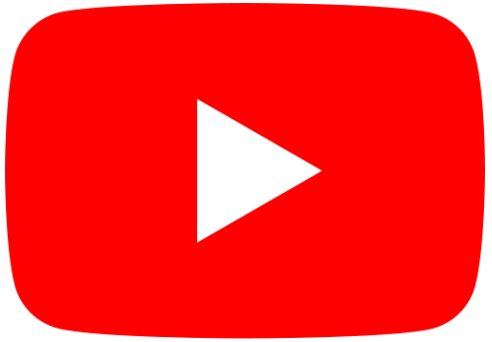
- A chatbot needs to understand the user's question in order to generate/retrieve a correct response.
- Example question:
  - "What are the best cafes around the Eiffel Tower?"
  - Chatbot needs to understand "Eiffel Tower" and "cafe" as locations and identify cafes within a certain distance of the Eiffel Tower.
- IE is useful in extracting such specific information from a pool of available data.





Social Media

- An example use case:
  - Extracting time-sensitive, frequently updated information, such as traffic updates and disaster relief efforts, based on tweets.
- NLP can help extract informative excerpts from social media text may help in decision making.



Extracting Data from Forms and Receipts

- Many banking apps have the feature to scan a check and deposit money directly into the user's account.
- Apps that scan bills and receipts are commonly used by individuals, small businesses, and larger enterprises.

IE Tasks

IE Tasks

- IE is a term that's used to refer to a range of different tasks of varying complexity.
- The overarching goal of IE is to extract "knowledge" from text.

To understand what these tasks are, consider the snippet from a *New York Times* article:

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.

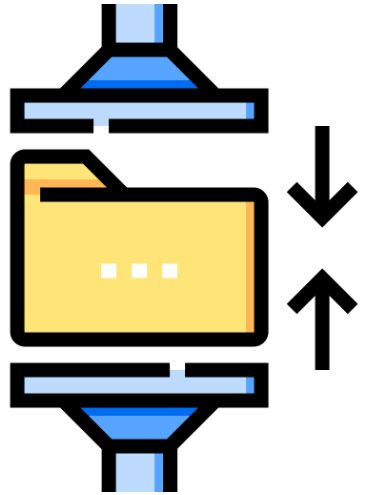
On Tuesday, Apple announced its plans for another major chunk of the money: It will buy back a further $75 billion in stock.

"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.

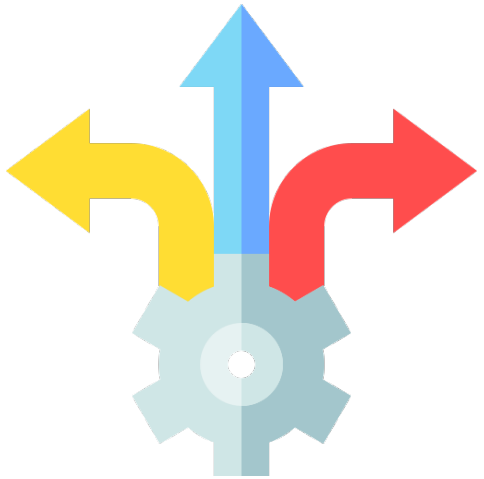
IE Tasks

- Keyword or Keyphrase Extraction (KPE)
  - Identifies article as being about "buyback" or "stock price"
- Named Entity Recognition (NER)
  - Identifies Apple as an organization and Luca Maestri as a person
- Named Entity Disambiguation and Linking
  - Recognizes that Apple refers to Apple, inc. And not some other company with the word "apple" in its name
- Relation Extraction
  - Extracts information that Luca Maestri is the finance chief of apple

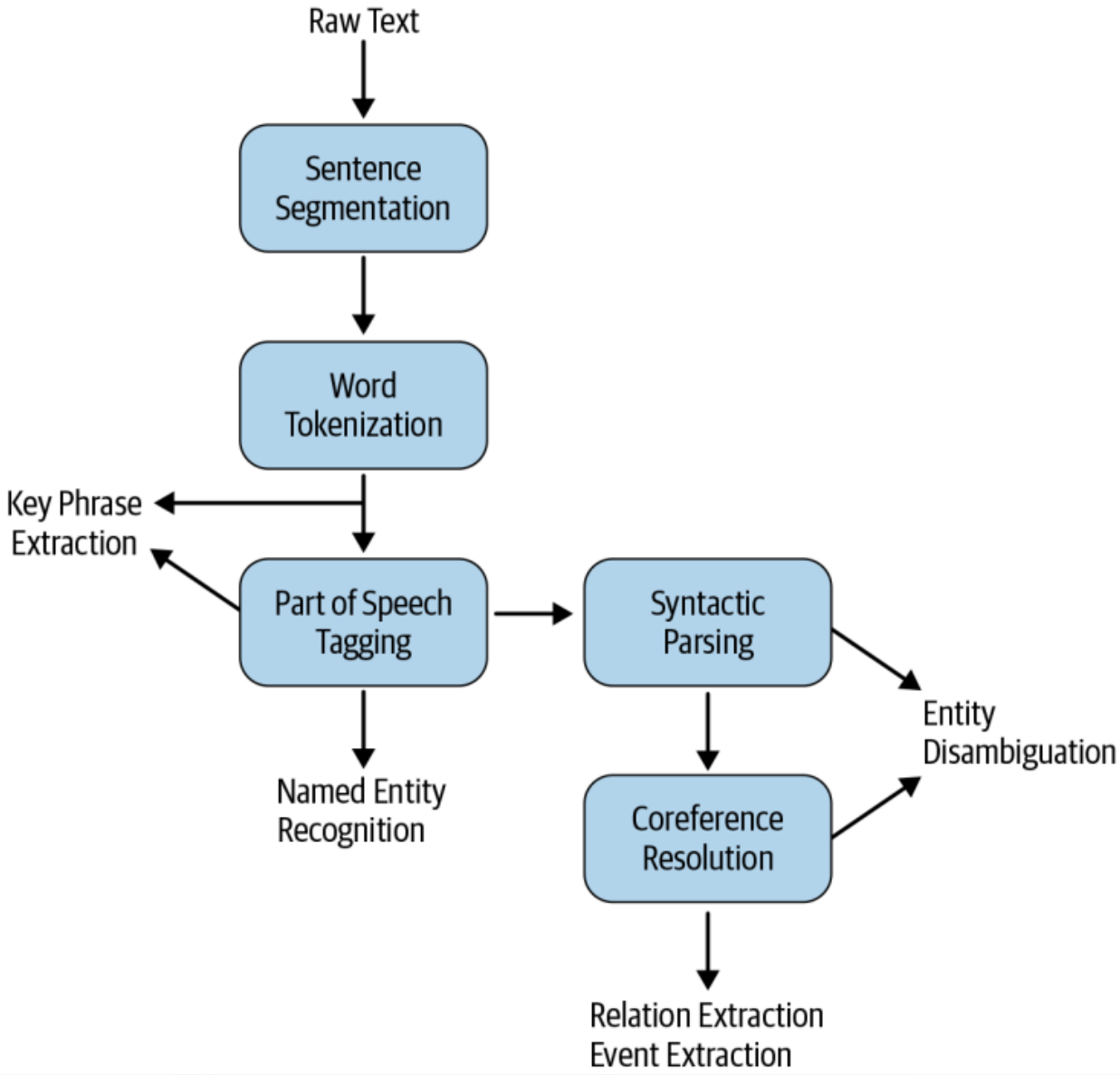


Approaches for IE

- Range of methods can be used
  - Rule-based
  - Machine learning (supervised, unsupervised, semi-supervised)
  - Deep learning
- IE in industry is generally implemented as a hybrid system incorporating rule-based and learning-based approaches.



General Pipeline of the IE Process



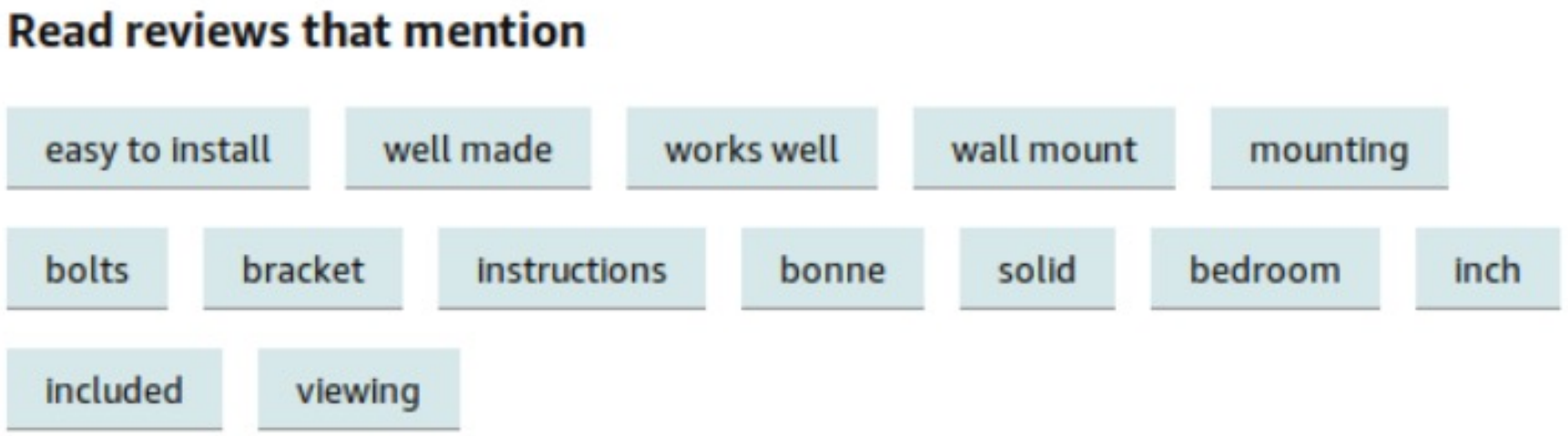
Keyphrase Extraction

Keyphrase Extraction

- The IE task concerned with extracting important words and phrases that capture the gist of the text from a given document.
- Useful for several downstream NLP tasks.
  - Such as search/information retrieval, automatic document tagging, recommendation systems, text summarization, etc.
- Two most commonly used methods:
  - **Supervised learning approaches** require labeled datasets, which are time- and cost-intensive to create.
  - **Unsupervised approaches** are more popular for KPE.

Keyphrase Extraction

- Amazon has a filtering feature for product reviews called "Read reviews that mention"
- Presents keywords or phrases that several people used in reviews to filter reviews



Coding Examples

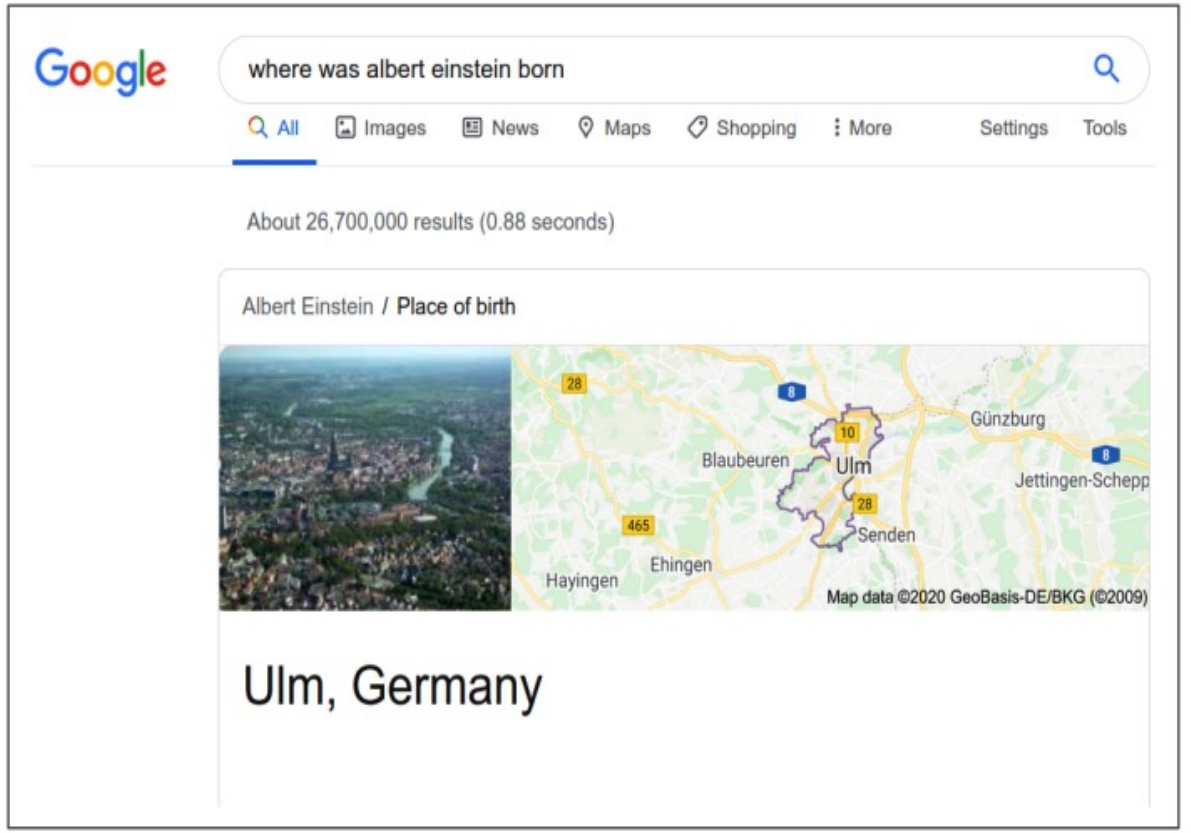
Practical Advice

- For some methods, the process of extracting potential n-grams is sensitive to document length.
  - Solution: Not use the full length, but instead try using the first M% and the last N% of the text.
- Since each keyphrase is independently ranked, we sometimes end up seeing overlapping keyphrases.
  - Solution: Use some similarity measure (e.g., cosine similarity) between the top-ranked keyphrases and choose the ones that are most dissimilar to one another.
- Seeing counterproductive patterns (e.g., a keyphrase that starts with a preposition when you don't want that) is another common problem.
  - Solution: Tweak the implementation code for the algorithm and explicitly encode information about such unwanted word patterns.
- Improper text extractions and document structure can affect the rest of KPE process.
  - Solution: Add some post-processing to extract key phrase list to create a final, meaningful list without noise.

Name Entity Recognition

Named Entities Recognition (NER)

To show “Ulm, Germany” for this query, the search engine needs to decipher that Albert Einstein is a **person** before going on to look for a place of birth:



Named Entities Recognition (NER) Concepts

- Named entity recognition (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.
- NER is an important step in the pipeline of several NLP applications involving information extraction.

How is NER Used?

- Healthcare
  - Improve patient care standards and reduce workloads by extracting essential information from lab reports
- Content classification
  - Surface content more easily and gain insights into trends by identifying the subjects and themes of blog posts and news articles
- Human resources
  - Speed up the hiring process by summarizing applicants' CVs; improve internal workflows by categorizing employee complaints and questions
- Customer support
  - Improve response times by categorizing user requests, complaints and questions and filtering by priority keywords

(1) Building an NER System: Rule-based

- Simple approach to building an NER system is to maintain a gazetteer of person/organization/location names.
  - To check whether a given word is a named entity, just do a lookup in the gazetteer.
  - This is a great way to start, especially when we don't have an existing NER system available.
- (Advanced) Rule-based NER goes beyond a lookup table and is based on a compiled list of patterns based on word tokens and POS (part of speech) tags.
  - Rules can be programmed to cover as many cases as possible to build a rule-based NER system
  - Stanford NLP's RegexNER and spaCy's EntityRuler provide functionalities to implement a rule-based NER system

(2) Building an NER System: ML

- Normal classifiers classify each word independently without considering the surrounding words in the sentence.
- However, NER is traditionally modeled as a sequence classification problem, where the entity prediction for the current word depends on the context (the entities of the previous and subsequent words).
  - For example, if the previous word was a person name, there's a higher probability that the current word is also a person name if it's a noun (e.g., first and last names).
- Conditional random fields (CRFs) is one of the popular sequence classifier training algorithms.

NER Using an Existing Library

- Many off-the-shelf libraries for NER are available.
- Popular libraries are Stanford NER, spaCy, and AllenNLP.
- We may run into two issues while using existing libraries
  - We may be using NER in a specific domain, and the pre-trained models may not capture the specific nature of our own domain.
  - Sometimes, we may want to add new categories to the NER system without having to collect a large dataset for all common categories.
- Many existing libraries can be customized.

Practical Advice

- Start with a pre-trained NER model and enhance it with heuristics, active learning, or both.
- NER is very sensitive to the format of its input. It is more accurate with well-formatted plain text than with, say, a PDF document from which plain text needs to be extracted first.
- NER is also very sensitive to the accuracy of the prior steps in its preprocessing pipeline. Some amount of pre-processing may be necessary before passing a piece of text into an NER model to extract entities.

POS Tagging

- POS tagging is a shallow NLP task that can be performed using simple rule-based methods or more complex machine learning techniques.
- The **aim** of POS tagging is to identify the syntactic role of each word in a sentence.
- The output of POS tagging is a sequence of tags, such as noun, verb, adjective, etc.

Results

NER

```
(S
  (PERSON Bill/NNP)
  (PERSON Gates/NNP)
  founded/VBD
  (ORGANIZATION Microsoft/NNP)
  Corp./NNP
  together/RB
  with/IN
  (PERSON Paul/NNP Allen/NNP)
  in/IN
  1975/CD
  ./.)
```

POS Tagging

```
[('Bill', 'NNP'),
 ('Gates', 'NNP'),
 ('founded', 'VBD'),
 ('Microsoft', 'NNP'),
 ('Corp.', 'NNP'),
 ('together', 'RB'),
 ('with', 'IN'),
 ('Paul', 'NNP'),
 ('Allen', 'NNP'),
 ('in', 'IN'),
 ('1975', 'CD'),
 ('.', '.')]
```

Did you notice any differences?

POS Tagging vs. NER

- Difference:
  - Objectives and output.
    - POS tagging is focused on identifying the syntactic role of words in a sentence, whereas NER is focused on identifying and classifying named entities in a text.
    - The output of POS tagging is a sequence of tags, whereas the output of NER is a set of entities along with their categories.
- Similarities:
  - Important NLP tasks used for information extraction.
  - Both tasks involve analyzing the text and assigning labels to each word based on their context.
  - Both tasks can be performed using simple rule-based methods or more complex machine learning techniques.

Take 10 minutes break…

Relation Extraction

Relationship Extraction

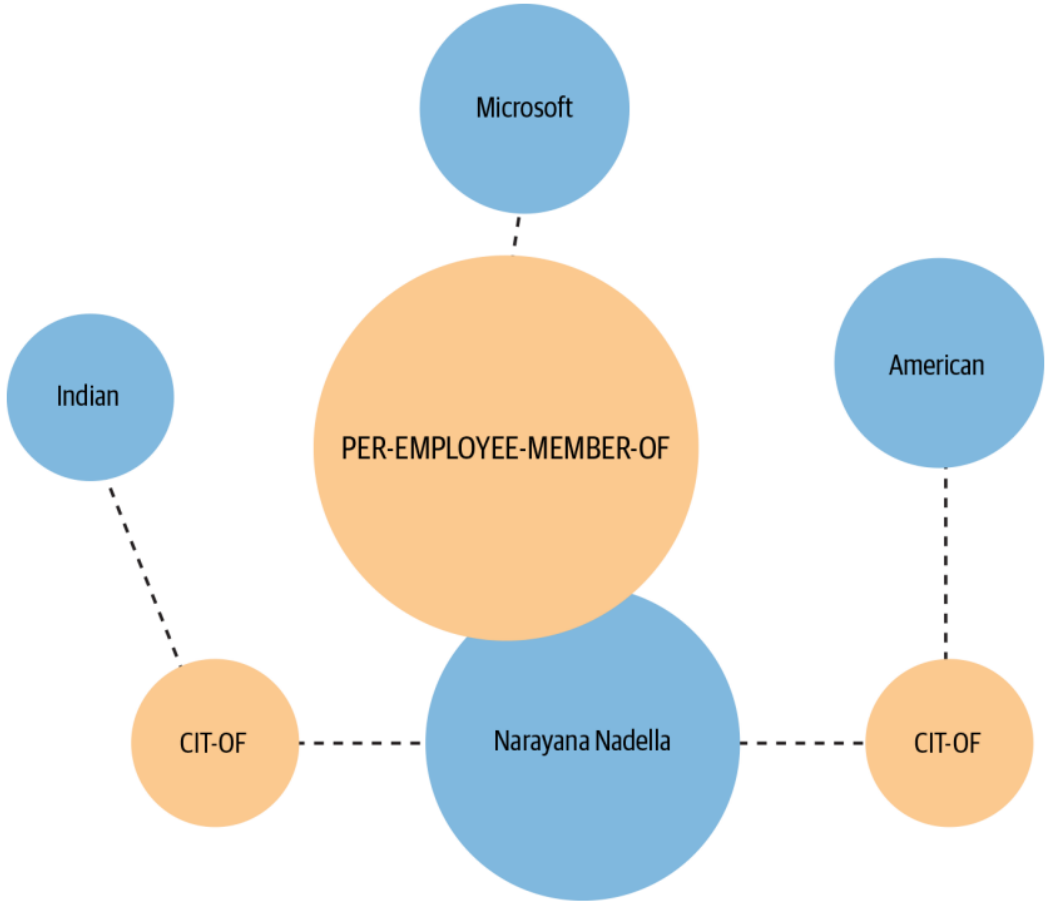
- Relationship extraction (RE) is the IE task that deals with extracting entities and relationships between them from text documents.
  - It's useful in improving search and developing question-answering system.
- An example of a working RE system by Rosette Text Analytics:

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.

- This 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.

An Example of A Working RE System

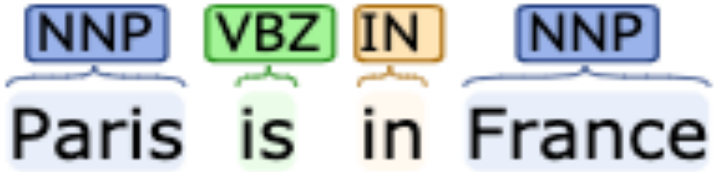
- What constitutes a “relation”?
- Relations can be specific to a given domain.
  - For example, in the medical domain, relations could include type of injury, location of injury, cause of injury, treatment of injury, etc.
  - In the financial domain, relations could mean something completely different.



(1) Rule-based RE (Hand-built Patterns)

- **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.
- We can also take into account the part-of-speech (POS) tags to enhance the precision.

Named entities in sentence



Part-of-speech tags in sentence

(1) Rule-based RE (Hand-built patterns)

- **Pros:**
  - Humans can create pattern which tend to have high precision
  - Can be tailored to specific domains
- **Cons:**
  - Human patterns are still often low-recall (too much variety in languages)
  - A lot of manual work to create all possible rules.
  - Have to create rules for every relation type. It could be challenging to create patterns to cover all possible relations within a domain.

(2) Supervised RE

- A common way to do supervised relation extraction is to train a stacked binary classifier (or a regular binary classifier) to determine if there is a specific relation between two entities.
- These classifiers take features about the text as input.
- Typical features are: context words, part-of-speech tags, dependency path between entities, NER tags, tokens, proximity distance between words, etc.

(2) Supervised RE

We could train and extract by:

1. Manually label the text data according to if a sentence is relevant or not for a specific relation type. E.g. for the "CEO" relation:
   - "Apple CEO Steve Jobs said to Bill Gates." is relevant
   - "Bob, Pie Enthusiast, said to Bill Gates." is not relevant
2. Manually label the relevant sentences as positive/negative if they are expressing the relation. E.g. "Apple CEO Steve Jobs said to Bill Gates.":
   - (Steve Jobs, CEO, Apple) is positive
   - (Bill Gates, CEO, Apple) is negative
3. Learn a binary classifier to determine if the sentence is relevant for the relation type
4. Learn a binary classifier on the relevant sentences to determine if the sentence expresses the relation or not
5. Use the classifiers to detect relations in new text data.

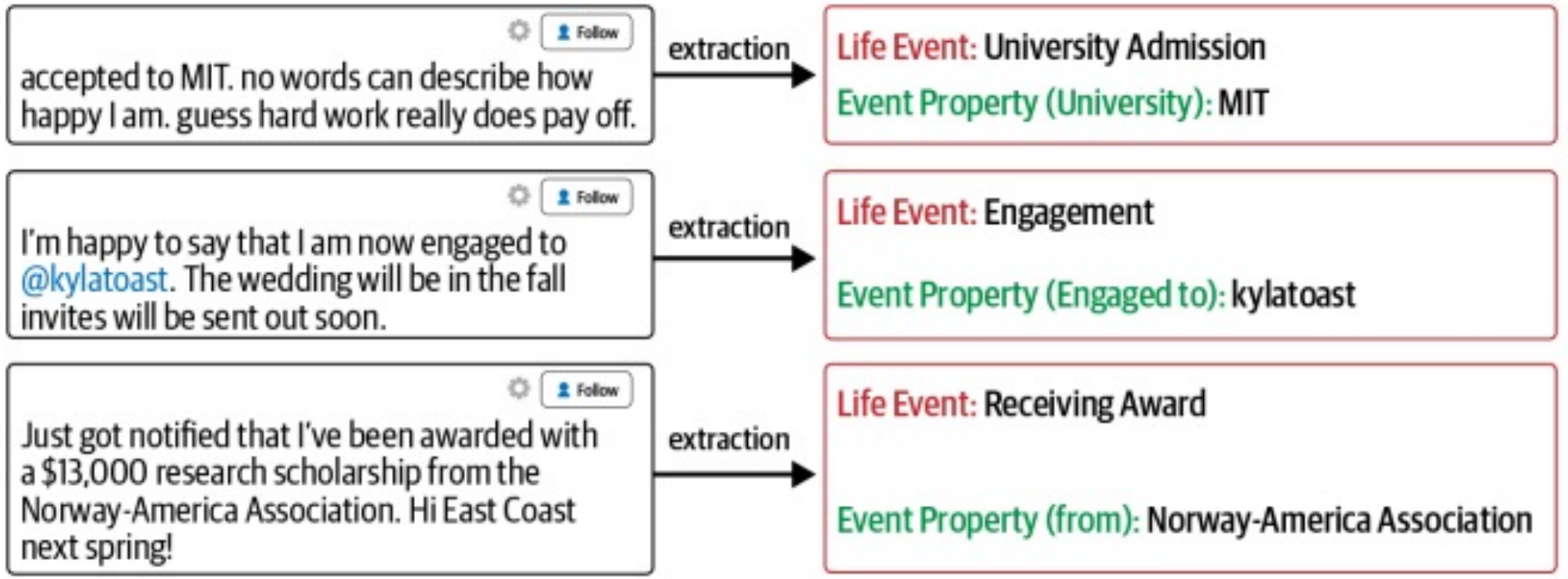
(2) Supervised RE

- **Pros:**
  - High quality supervision (ensuring that the relations that are extracted are relevant)
  - We have explicit negative examples.
- **Cons:**
  - Expensive to label examples
  - Expensive/difficult to add new relations (need to train a new classifier)
  - Does not generalize well to new domains
  - It is only feasible for a small set of relation types.

Advanced IE Tasks

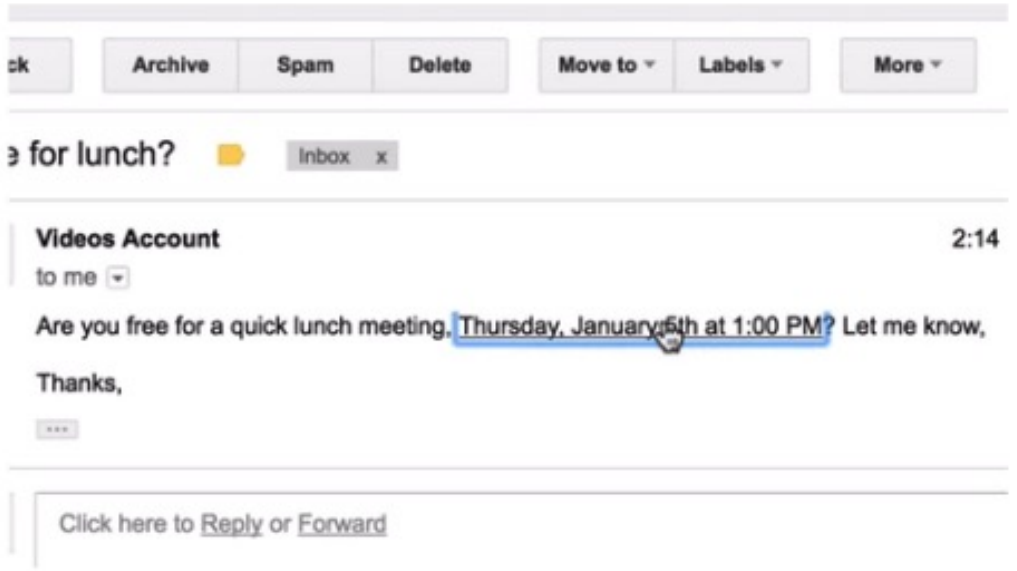
1. Event Extraction

- **Event extraction** is the IE task that deals with identifying and extracting events from text data.
- Events can be anything that happens at a certain point in time.
  - Such as meetings, increase in fuel prices or life events like birth, marriage, and demise.



2. Temporal Information Extraction

- Temporal IE and normalization is an IE task that involves extracting date and time information from text data and converting it into a standard form.
- Applications: email calendar utilities.
- Methods: handcrafted patterns, supervised sequence labeling techniques.
- Existing libraries: duckling

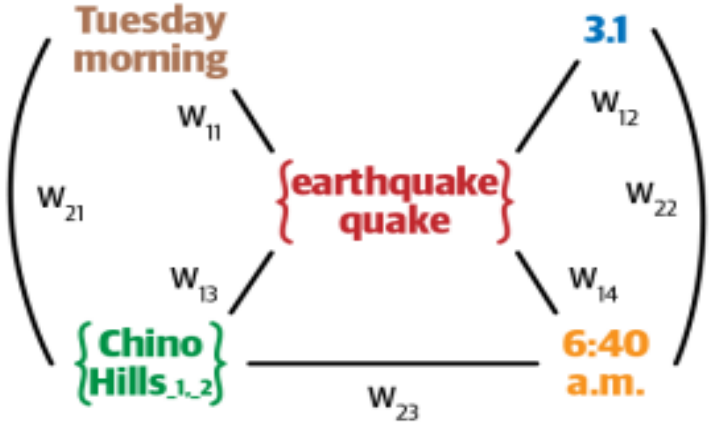


3. Template Filling

- Template filling is an IE task that models text generation as a slot-filling problem.
- Use cases: automate report generation for cases like weather forecasts and financial reports.

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




QUIZ

Q1. What is the main objective of Keyword or Keyphrase Extraction (KPE)?

A. Identifying people and organizations mentioned in a text
B. Identifying the tone or sentiment of a text
C. Identifying the main information from a text
D. Identifying the location where a text was written

Answer: C

Q2. What does Named Entity Recognition (NER) aim to identify in a text?

A. The main topics or themes of text
B. The tone or sentiment of the text
C. The location where the text was written
D. People, organizations, locations, and other named entities mentioned in the text

Answer: D

Q3. POS tagging is focused on identifying the syntactic role of words in a sentence.

A. True
B. False

Answer: A

Q4. _____ is a package and designed to parse text and get structured data, it can process the natural language text data to extract temporal events.

A. NER
B. NEL
C. NED
D. Duckling

Answer: D

Q5. We can use both supervised and unsupervised methods for Keyphrase Extraction.

A. True
B. False

Answer: A


SPRING
BREAK!

Exercises using Google Colab