NLP Applications II

Introduction to Information Extraction

Information Extraction

- **Information extraction (IE) systems**
  - Find and understand the relevant parts of texts
  - Gather information from many pieces of text
  - Produce a structured representation of relevant information:
    - *relations* (in the database sense), a.k.a.,
    - a *knowledge base*
- **Goals**:
  - Organize information so that it is useful to people
  - Put information in a semantically precise form that allows further inferences to be made by computer algorithms

Information Extraction (IE)

- IE systems extract clear, factual information
  - Roughly: *Who did what to whom when?*
- E.g.,
  - Gathering earnings, profits, board members, headquarters, etc. from company reports
    - *The headquarters of BHP Billiton Limited, and the global headquarters of the combined BHP Billiton Group, are located in Melbourne, Australia.*
    - `headquarters`(“BHP Biliton Limited”, “Melbourne, Australia”)
- Learn drug-disease interactions from medical research literature
  - *Igualmente la cefalea del paciente a mejorado al disminuir la dosis de nitroglicerina*

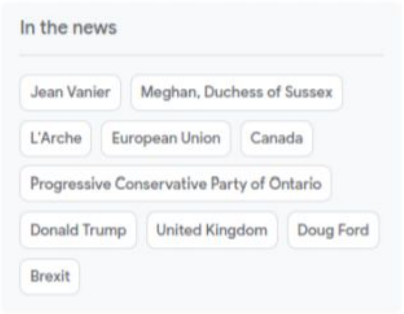
`causeBy`

Entities, Relations, Events

Much of information extraction can be described in terms of **entities**, **relations**, and **events**.

- **Entities** are uniquely specified objects in the world.
  - people (JEFF MANGUM), places (ATHENS, GEORGIA), organizations (MERGE RECORDS), and times (FEBRUARY 10, 1998).
- **Relations** include a **predicate** and **two arguments**:
  - `CAPITAL`(GEORGIA, ATLANTA).
- **Events** involve multiple typed arguments defined by a **schema**.
  - Events often refer to time-delimited occurrences: **weddings**, **protests**, **purchases**, **terrorist attacks**.

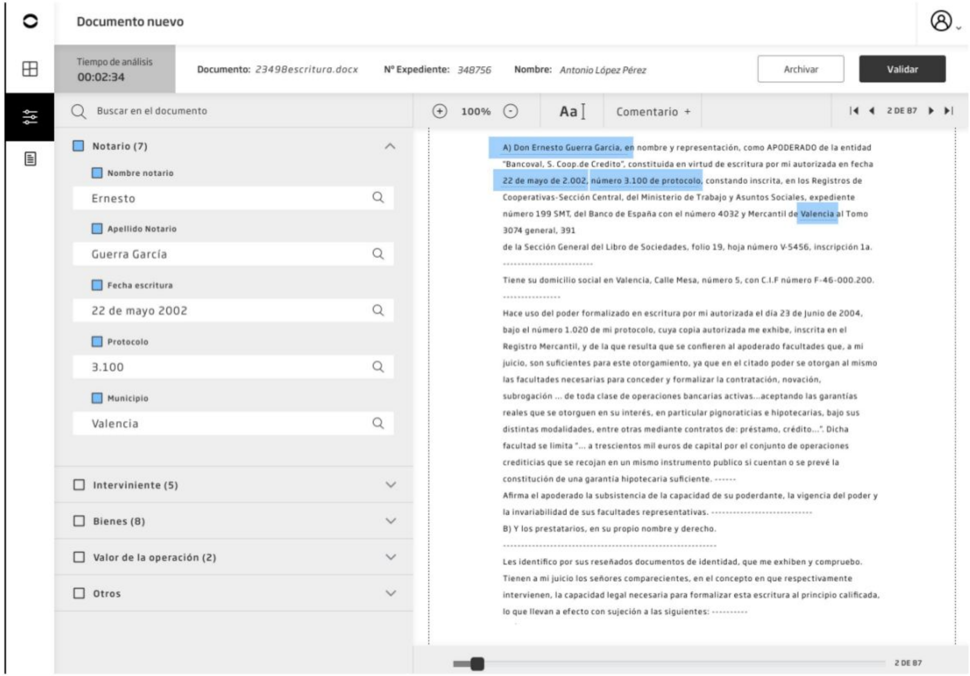
IE in real-world applications



- Tagging news and other content
  - Entities and events can be indexed.
  - Useful for search engines and recommender systems
- Chatbots
  - NER is critical part in the NLU component
  - "*What are the best cafes around Eiffel Tower?*" → need to know they are **locations**.
- Applications in social media
  - Extracting time-sensitive, frequently updated information, such as traffic updates and disaster relief efforts, based on tweets.
  - Sentiment analysis, opinion mining, fake news, content filtering…

IE in real-world applications

- Extracting data from legal documents
  - Automatize manual effort (information storage) in the back-office department
  - OCR needed most of the time
  - Distant supervision



IE tasks

April 30, 2019

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

April 30, 2019

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.

Keyword/phrase Extraction (KPE)

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

April 30, 2019

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.

Keyword/phrase Extraction (KPE)

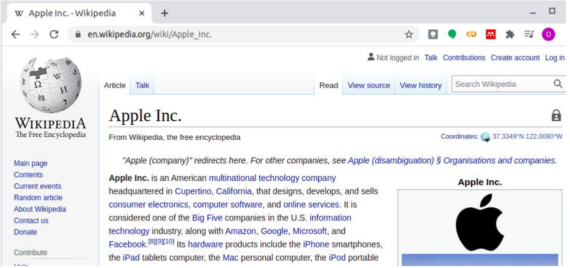
On Tuesday, Apple announced its plans for another major chunk of the money: It will buy back a further $75 billion in stock.

Named Entity Recognition (NER)
Apple - ORG
Luca Maestri - PER

"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



Named Entity Disambiguation

April 30, 2019

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.

Keyword/phrase Extraction (KPE)

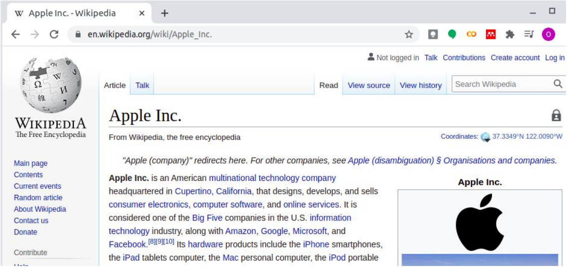
On Tuesday, Apple announced its plans for another major chunk of the money: It will buy back a further $75 billion in stock.

Named Entity Recognition (NER)
Apple - ORG
Luca Maestri - PER

“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



Named Entity Disambiguation

April 30, 2019

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.

Keyword/phrase Extraction (KPE)

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

Relation Extraction (RE)
CFO(*Luca Maestri*, *Apple inc.*)

Named Entity Recognition (NER)
Apple - ORG
Luca Maestri - PER

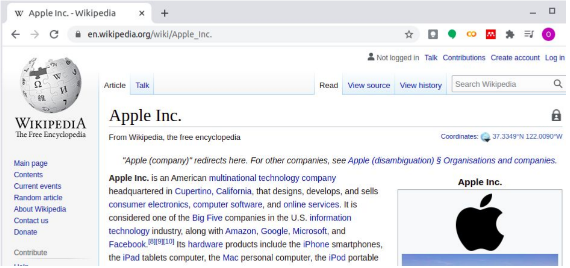
“Our first priority is always looking after the business and making sure we continue to grow and invest,” Luca Maestri, Apple’s finance chief, said in an interview. “If there is excess cash, then obviously we want to return it to investors.”

CFO

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

Named Entity Disambiguation



April 30, 2019

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.

Keyword/phrase Extraction (KPE)

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

Relation Extraction (RE)
CFO(*Luca Maestri, Apple inc.*)

Named Entity Recognition (NER)
Apple - ORG
Luca Maestri - PER

"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."

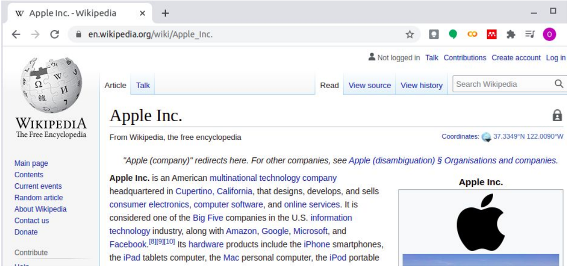
CFO

Event Extraction (RE)
Event: "Apple buys back stock"

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

Named Entity Disambiguation



April 30, 2019

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."

CFO

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.

Keyword/phrase Extraction (KPE)

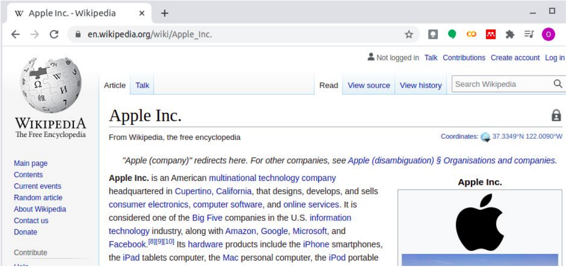
Named Entity Recognition (NER)
Apple - ORG
Luca Maestri - PER

Relation Extraction (RE)
CFO(*Luca Maestri*, *Apple inc.*)

Event Extraction (RE)
Event: "Apple buys back stock"

Temporal Information Extraction

IE tasks



Named Entity Disambiguation

April 30, 2019

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.

Keyword/phrase Extraction (KPE)

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

Named Entity Recognition (NER)
Apple - ORG
Luca Maestri - PER

“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.”

CFO

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.

Relation Extraction (RE)
CFO(*Luca Maestri*, *Apple inc.*)

Event Extraction (RE)
Event: “Apple buys back stock”

Temporal Information Extraction

Template filling
Event_type: buyback
Buyer: Apple Inc.
Theme: $100 billin of its stock

Keyword and phrase extraction

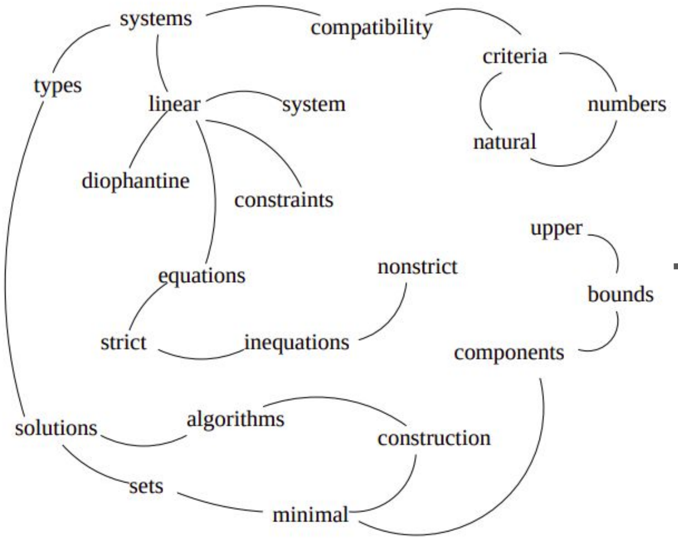
- Extract the important words and phrases that capture the main ideas of the text.
- Two common approaches
  - **Supervised learning** approaches require corpora with **texts and their respective keyphrases** and use engineered features or DL techniques.
  - Unsupervised way is to represent the words and phrases in a text as nodes in a **weighted graph** where the weight indicates the importance of that keyphrase.
    - **TextRank**: https://www.aclweb.org/anthology/W04-3252/
    - **SGRank**: https://www.aclweb.org/anthology/S15-1013.pdf
- Available implementations
  - Textacy: https://github.com/chartbeat-labs/textacy

TextRank

1. Identify **words or phrases** of text and add them as **vertices** in the graph.
2. Identify relations (**edges**) that connect such text units. Edges can be directed or undirected, weighted or unweighted.
   - **Co-occurrence** in a context-window of 2-10 words
3. Iterate the graph-based ranking algorithm until convergence (e.g. pageRank)
4. Sort vertices based on their final score. Use the values attached to each vertex for ranking/selection decisions.

TextRank

Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types systems and systems of mixed types.



Keywords assigned by TextRank:
linear constraints; linear diophantine equations; natural numbers; nonstrict inequations; strict inequations; upper bounds

Keywords assigned by human annotators:
linear constraints; linear diophantine equations; minimal generating sets; non−strict inequations; set of natural numbers; strict inequations; upper bounds

Keyword and phrase extraction

Input text about history of NLP

```
from textacy import *
import textacy.ke

mytext = open(“nlphistory.txt”).read()
en = textacy.load_spacy_lang("en_core_web_sm", disable=("parser",))
doc = textacy.make_spacy_doc(mytext, lang=en)

print("Textrank output: ", [kps for kps, weights in
textacy.ke.textrank(doc, normalize="lemma",  topn=5)])

print("SGRank output: ", [kps for kps, weights in
textacy.ke.sgrank(mydoc, n_keyterms=5)])

Output:
Textrank output:  ['successful natural language processing system',
'statistical machine translation system', 'natural language system',
'statistical natural language processing', 'natural language task']

SGRank output:  ['natural language processing system','statistical machine
translation', 'research', 'late 1980', 'early']
```

Output of textrank

Named Entity Recognition

- NER refers to the IE task of identifying the entities in a document
  - NER visualizer: https://explosion.ai/demos/displacy-ent
- Rule-based system do not generalize well (high precision, though):
  - https://spacy.io/usage/rule-based-matching (rule-based matching)
  - https://explosion.ai/demos/matcher

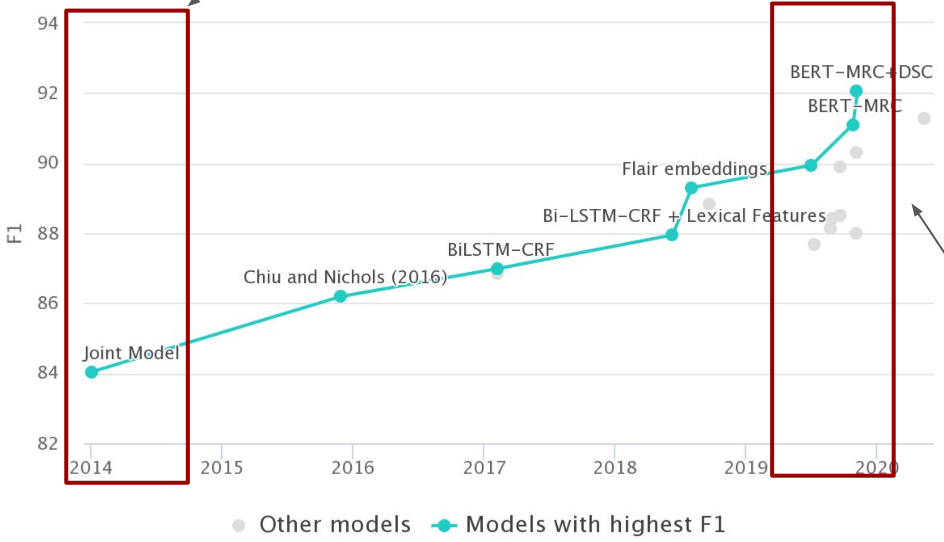
```
{(([{ ner:PERSON }]) /was/ /born/ /on/ ([ { ner:DATE } ])) =>"DATE_OF_BIRTH" }
```

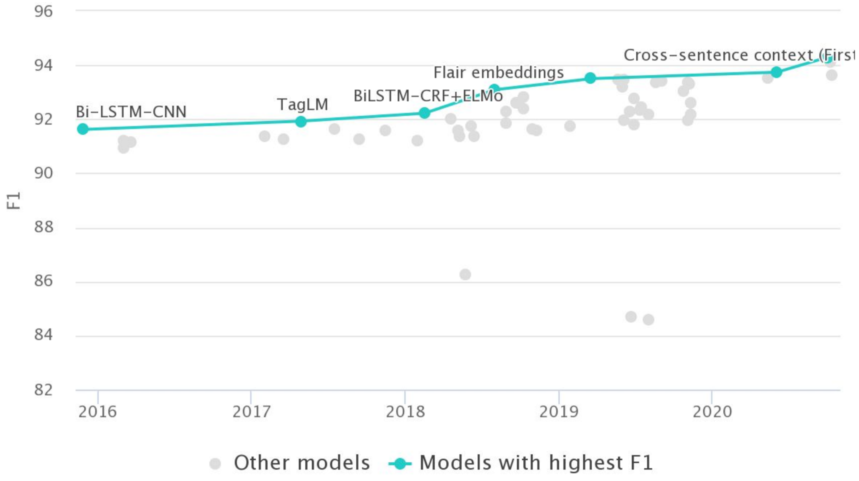
- ML/DL cast NER as a **sequence labeling** problem
  - Feature engineered ML: Conditional Random Fields (CRF) + BeamSearch/Viterbi
  - Deep Learning set the state of the art: LSTM → CNN → Transformers (BERT)

NER Leaderboard

Until DL, the NER was stacked

ontonotes v5.0





conll-2003

Impressive boost with transformer based models

NER using an existing library

- NER has been well researched over the past few decades.
- Exists off-the-shelf libraries that can be used to incorporate a pre-trained NER model into a software product
  - Stanford NER
  - spaCy
  - AllenNLP

```python
import spacy

nlp = spacy.load("en_core_web_lg")

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

doc = nlp(text_from_fig)

for ent in doc.ents:
    if ent.text:
     print(ent.text, "\t", ent.label_)
```

Encoding labels for sequence

IO encoding

Foreign	ORG
Ministry	ORG
spokesman	O
Shen	PER
Guofang	PER
told	O
Reuters	ORG
:	:

IOB encoding

Foreign	B-ORG
Ministry	I-ORG
spokesman	O
Shen	B-PER
Guofang	I-PER
told	O
Reuters	B-ORG
:	:

Foreign
Ministry
spokesman
Shen
Guofang
told
Reuters
:

Named Entity Disambiguation / Linking

- NEL: Assigning a unique identity to entities mentioned in the text.
  - NER and NED together are known as named entity linking (NEL)
- Requires knowledge of several IE tasks beyond what we’ve seen with NER and KPE.
  - “Lincoln drives a Lincoln Aviator and lives on Lincoln Way.”
- Some NLP applications need NEL
  - Question Answering
  - Constructing large knowledge bases of connected events and entities such as the Google **Knowledge Graph**

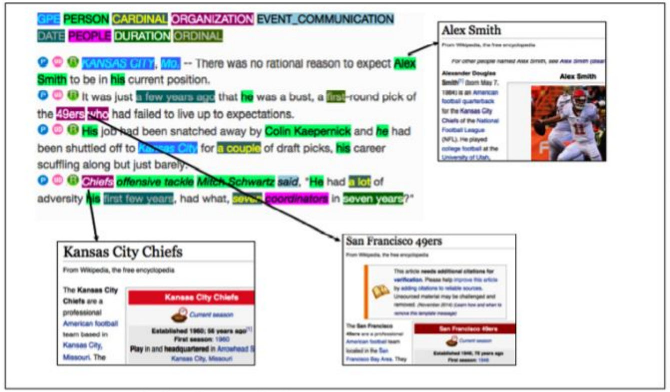


Figure 5-8. Entity linking by IBM [33]

Named Entity Disambiguation / Linking

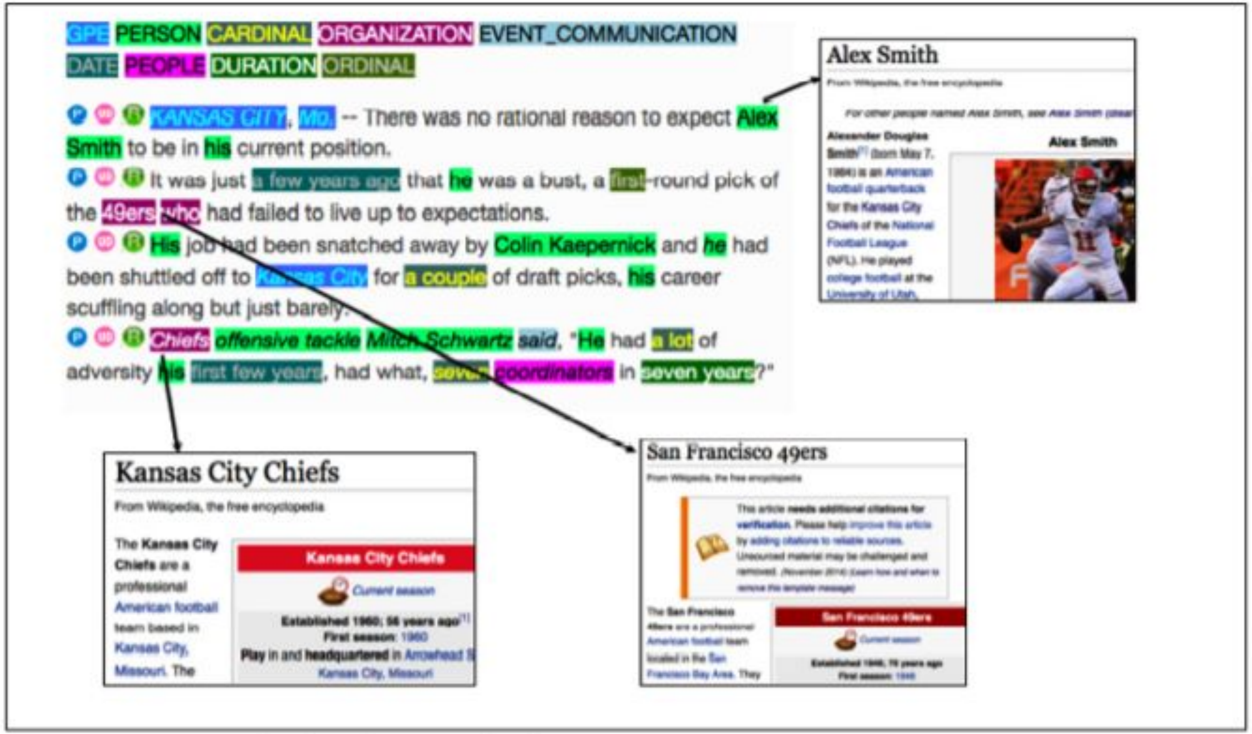


Figure 5-8. Entity linking by IBM [33]

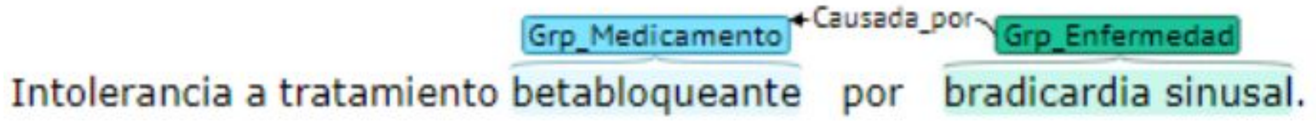
NEL vs NER

So, how do we build an IE system for performing NEL?

- **Relies on context**. Just as NER identifies entities and their spans using contextual information encoded by a range of features.
- It requires going **beyond POS tagging** in terms of the NLP pre-processing
- Needs some form of parsing to identify linguistic items like **subject, verb, and object**.
- Need **coreference resolution** to resolve and link multiple references to the same entity.
    - Albert Einstein = the scientist = Einstein
- Combines **local** and **global** information:
    - **Local**: Modeled as a **supervised ML/DL problem** (similar to NER)
    - **Global**: Collective entity linking, use graph-theoretic approaches over KB (e.g) to measure **entity coherence**
        1. California, Oregon, Washington
        2. Baltimore, Washington, Philadelphia
        3. Washington, Adams, Jefferson

Relation Extraction

- **Relation extraction** (RE) is a task that deals with extracting entities and relationships between them from text documents.
  - Knowledge based population, Advanced Search, Question Answering
  - Relations are specific to a given domain
    - Medical domain: Adverse Drug Reaction



Relation Extraction

- Imagine we're working at a company that mines tons of news articles to derive financial insights.
- To be able to do such analysis on thousands of news texts every day, we would need a constantly **updated knowledge base** that **connects** different **people**, **organizations**, and **events** based on the news content.
  - KPE, NER, NEL are helpful, but we need more.
  - "Connect" → extract relationship between the entities: `(Luca Maestri, finance chief, Apple)`.

Approaches to RE

- Requires deeper knowledge of language processing as compared than NER or KPE.
- **Hand-built patterns** (regexp) have high precision, but low recall
- **Supervised classification** (similar to text classification)
  - Binary classification: whether two entities in text are related
  - Multiclass classification: if they are related, what is the relation type (from a predefined set)
- Typical **learning features** are:
  - Words around and between the two entities
  - Syntactic structures (e.g NP VP NP)
  - Entity types
- **Deep Learning** beyond simple text classification or sequence labelling
  - Need to mark arguments of the relation

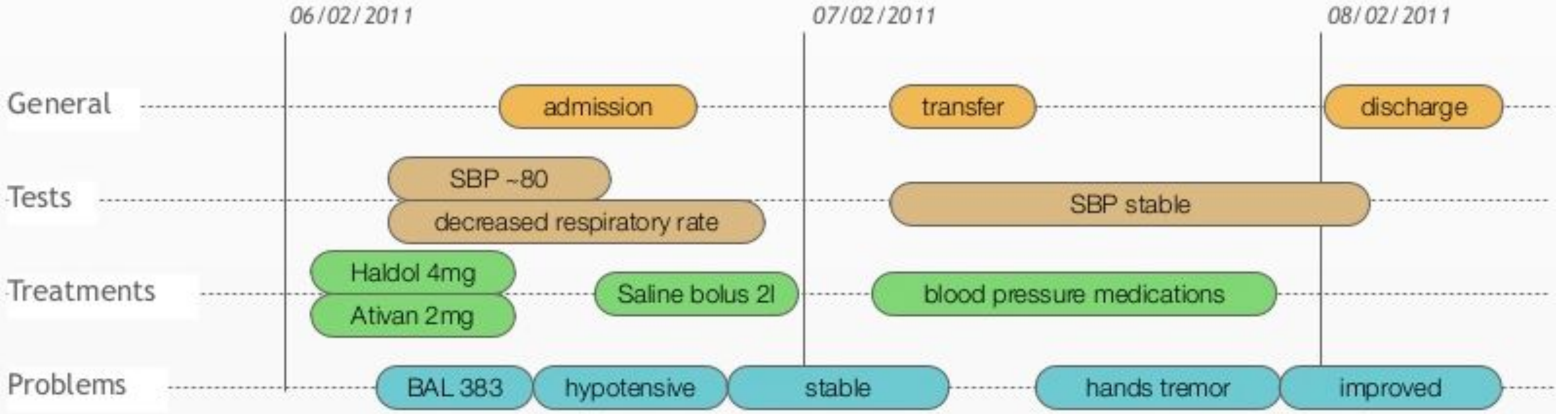
Approaches to RE

- Domain specific relationship set, and need **large amounts of annotated data**
- **Bootstrapping**
  - Starting with a small set of seed patterns and generalizing by learning new patterns based on the sentences extracted using these seed pattern
- **Distant supervision**
  - Use large databases (Wikipedia, Freebase) to first collect thousands of examples of many relations, and use ML to train a supervised classification.
- **Open IE** (unsupervised RE): (https://demo.allennlp.org/)
  - Extract relation without any list of relations and training instances
    - Relations are <verb, argument1, argument2> tuples

Temporal Information Extraction

- **Temporal expressions**: phrases denoting a temporal entity such as an interval or time point.
  - 01/05/2014, March 15, the next week, Saturday, at that time, yesterday, 5 o’clock, 3 days, every 4 hours
- **Events**: phrases denoting eventuality and states
  - Inflected verbs and nouns: spoken, deliver, will be published
- **Links**: tempora relation between two phrases
  - BEFORE, AFTER, INCLUDES, ENDS, DURING, BEGINS
- **Timelines**: order events in the temporal axis

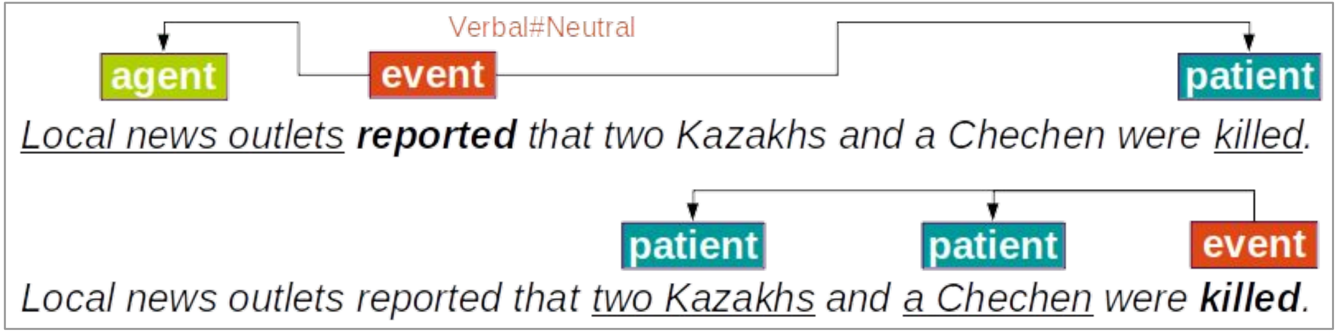
ADMISSION DATE: 2011-02-06;
DISCHARGE DATE: 2011-02-08;
HISTORY OF PRESENT ILLNESS: Mr. Pohl is a 53 - year-old male with history of alcohol use and hypertension. Blood alcohol level was 383. Agitated in emergency room requiring 4 leather restraints, received 5 mg of Haldol, 2 mg of Ativan. He became hypotensive in the emergency room with a systolic blood pressure in the 80 's and had decreased respiratory rate. He received a normal saline bolus of 2 litres of good blood pressure response. The patient was then admitted to the medical Intensive Care Unit for observation and then transferred to our service on medicine when the blood pressures remained stable overnight...



Event Extraction

- 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 task that deals with identifying and extracting events from text data.
- Usually, a **schema** (e.g.participants, time span…) is provided for each event type (e.g. terrorist attack, chemical reaction)
  - Indicate all the possible properties of the event (e.g. `perpetrator`, `victim`, `target`, `instrument`).
  - The system is then required to fill in as many of these properties as possible.

Examples of event extraction



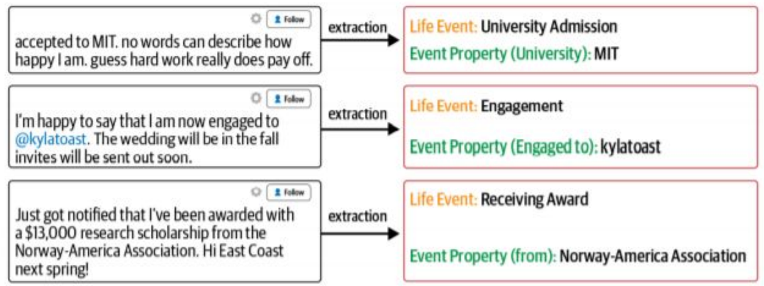


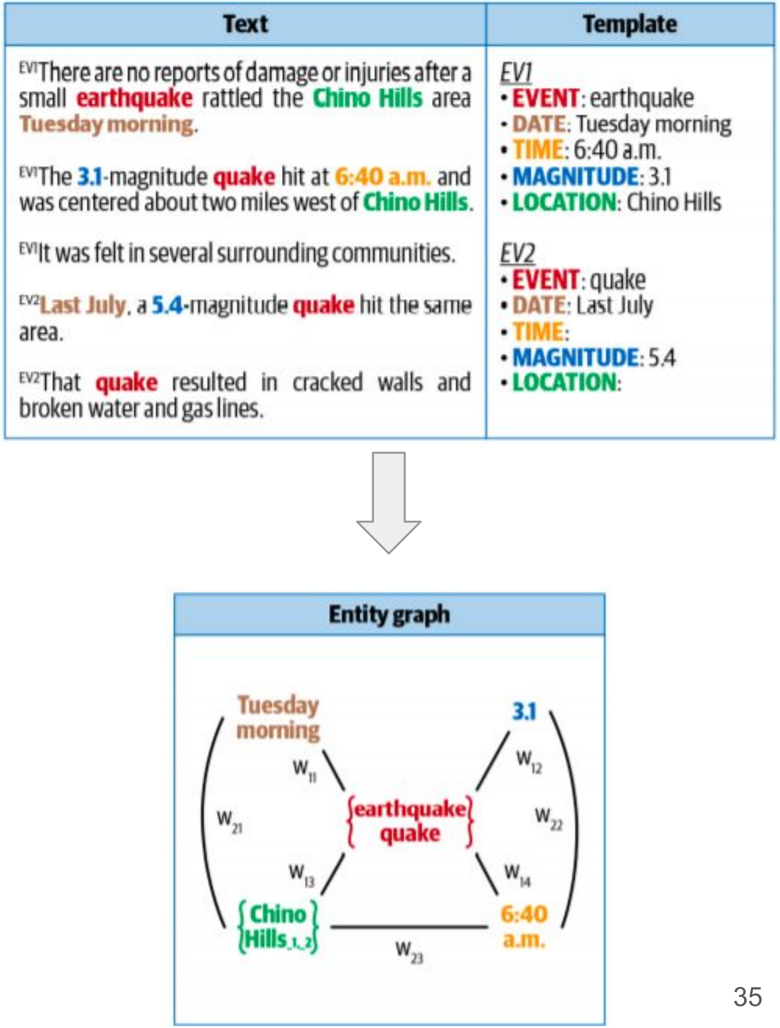
Figure 1: Examples of users mentioning personal life events on Twitter.

Figure 5-15. Examples of extracting life events from Twitter data [50]

[EVENT Citing] high fuel prices, United Airlines [EVENT said] Friday it has [EVENT increased] fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately [EVENT matched] [EVENT the move], spokesman Tim Wagner [EVENT said]. United, a unit of UAL Corp., [EVENT said] [EVENT the increase] took effect Thursday and [EVENT applies] to most routes where it [EVENT competes] against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

Template filling

- Very close to event extraction.
- Need to fill the slot of predefined templates (associated to event type),
- 1) Identify whether a template is present in a given sentence (template recognition)
    a. Text classification
    b. Trigger identification (sequence labelling)
- 2) Identify slot fillers for that template (slot-filling)
    a. Separate classifier for each slot
    b. Sequence labeller



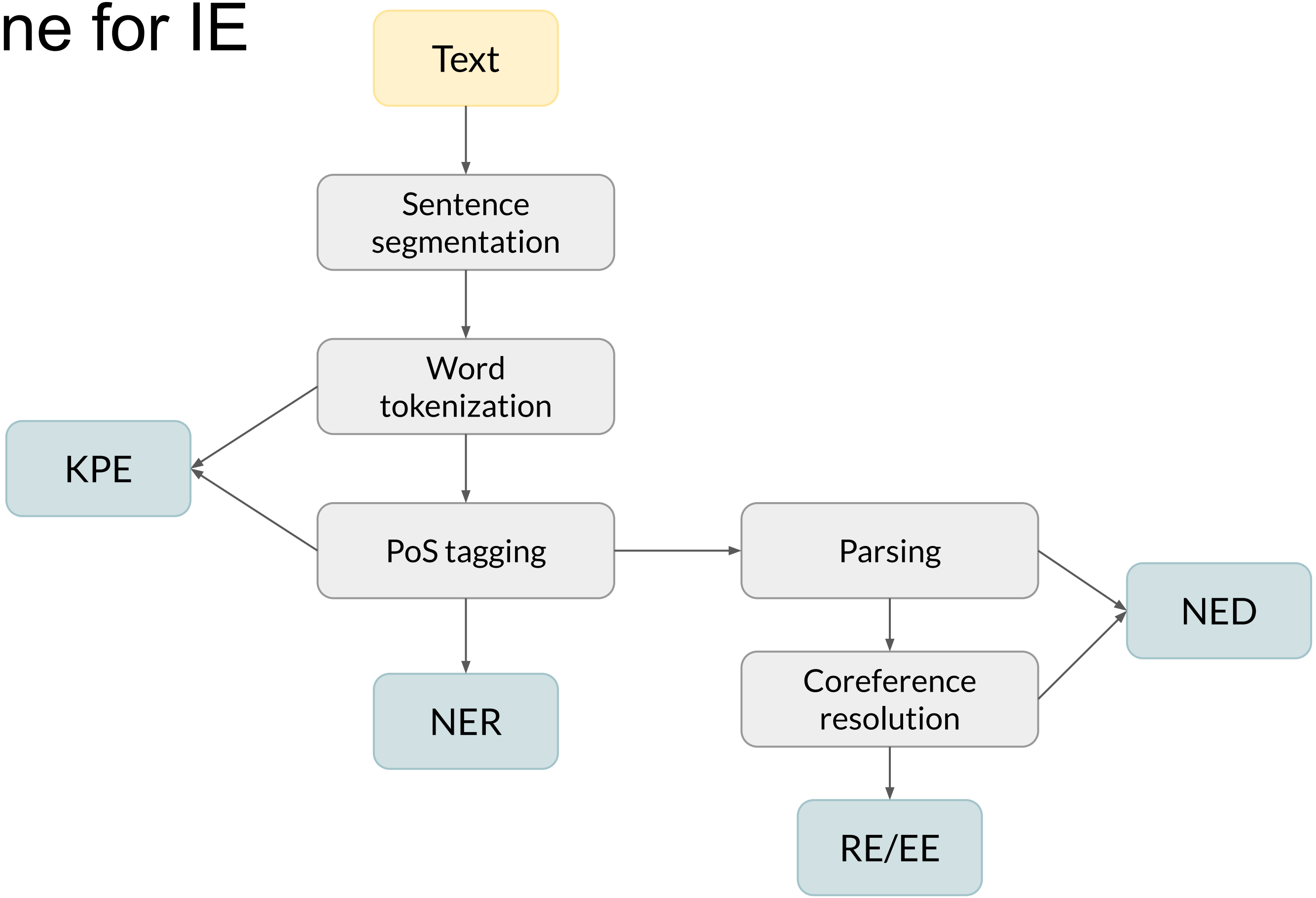
Sequence tagging

- Many of the task we review today are approached with sequence labelling.
  - Instead of having IOB labels for NER (`ORG, PER, LOC, MISC`) we can have any other annotation schema
    - `[O O O B-EVENT O O O]`
- There are great libraries to build sequence labeler in few lines of code
  - Scikit-learn, Tensorflow, Pytorch, spaCy

```
self.model = tf.keras.models.Sequential([
   tf.keras.layers.Input(shape=(None,), dtype='int32', name='word_ids'),
   tf.keras.layers.Embedding(self.config.nwords, self.config.dim_word,
                             mask_zero=True,
                             trainable=self.config.train_embeddings)

   tf.keras.layers.Bidirectional(tf.keras.layers.LSTM(units=self.config.hidden_size_lstm,
                                                      return_sequences=True)),
   tf.keras.layers.Dropout(self.config.dropout),
   tf.keras.layers.Dense(self.config.number_labels, activation='softmax')
  ]
)
```

A pipeline for IE



Evaluation

Metric	Description	Applications
Accuracy	Used when the output variable is categorical or discrete. It denotes the fraction of times the model makes correct predictions as compared to the total predictions it makes.	Mainly used in classification tasks, such as sentiment classification (multiclass), natural language inference (binary), paraphrase detection (binary), etc.
Precision	Shows how precise or exact the model's predictions are, i.e., given all the positive (the class we care about) cases, how many can the model classify correctly?	Used in various classification tasks, especially in cases where mistakes in a positive class are more costly than mistakes in a negative class, e.g., disease predictions in healthcare.
Recall	Recall is complementary to precision. It captures how well the model can recall positive class, i.e., given all the positive predictions it makes, how many of them are indeed positive?	Used in classification tasks, especially where retrieving positive results is more important, e.g., e-commerce search and other information-retrieval tasks.
F1 score	Combines precision and recall to give a single metric, which also captures the trade-off between precision and recall, i.e., completeness and exactness. F1 is defined as (2 × Precision × Recall) / (Precision + Recall).	Used simultaneously with accuracy in most of the classification tasks. It is also used in sequence-labeling tasks, such as entity extraction, retrieval-based questions answering, etc.
AUC	Captures the count of positive predictions that are correct versus the count of positive predictions that are incorrect as we vary the threshold for prediction.	Used to measure the quality of a model independent of the prediction threshold. It is used to find the optimal prediction threshold for a classification task.

Source: http://www.practicalnlp.ai/

Further Reading

- Jurafsky, Daniel and James H. Martin. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition**. 3rd Edition
  - Book https://web.stanford.edu/~jurafsky/slp3/
  - Chapter 17. Information Extraction: https://web.stanford.edu/~jurafsky/slp3/17.pdf
- Jacob Eisenstein. **Introduction to Natural Language Processing**.
  - Chapter 17. Information Extraction https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf

Wrap up

- We introduce information extraction as set of approaches to extract relevant information in real-world scenarios.
- IE task require resource beyond annotated text such as domain specific knowledge.
- Thus, it is good idea to build systems with existing pre-trained models.
- Reasonable way to approach many task is as sequence labeling
  - Take a look on CRFs, LSTMs and Transformers