BITS Pilani
Pilani Campus

Natural Language Processing Applications

Dr. Chetana Gavankar, Ph.D,
IIT Bombay-Monash University Australia
Chetana.gavankar@pilani.bits-pilani.ac.in

BITS Pilani
Pilani Campus

Session 16: Endsem Review Part 2
Date – 31st March 2024
Time – 1.40 pm to 3.40 pm

These slides are prepared by the instructor, with grateful acknowledgement of Prof. **Philipp Koehn ,** Prof. Raymond J. Mooney, Prof. Jurafsky, Abigail See, Matthew Lamm and many others who made their course materials freely available online.

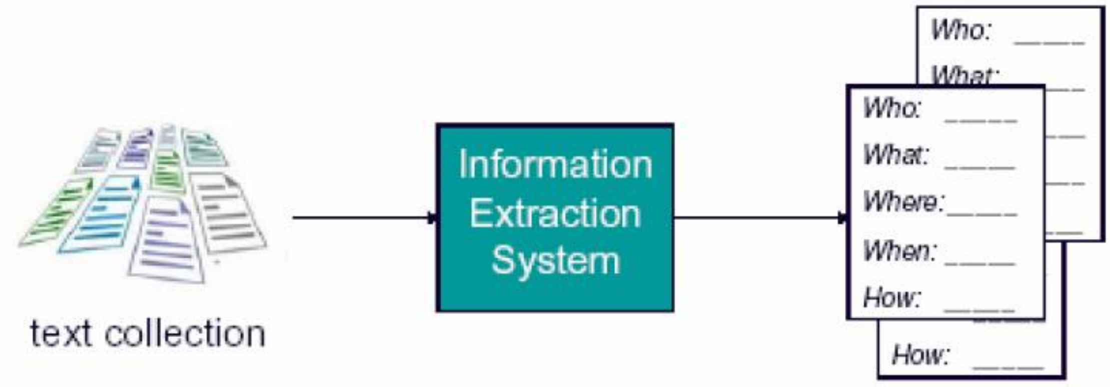
Agenda

- **Information Extraction Review**
- Named Entity Recognition
- Relation Extraction
- Temporal and Event Recognition

- **Sentiment Analysis Review**

Information Extraction (IE)

- Identify specific pieces of information (data) in an unstructured or semi-structured text
- Transform unstructured information in a corpus of texts or web pages into a structured database (or templates)
- Applied to various types of text, e.g.
  - Newspaper articles
  - Scientific articles
  - Web pages



Source: J. Choi, CSE842, MSU

A Typical IE Processing Pipeline

Named Entity Recognition (NER) & Shallow Parsing

↓

Reference Resolution

↓

Relation Detection & Classification

↓

Event Detection & Classification

↓

Template Filling

What is Information Extraction

As a task: ***Filling slots in a database from sub-segments of text.***

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

IE

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

Adapted from slide by William Cohen

What is Information Extraction?

As a family of techniques:

Information Extraction = segmentation + classification + association

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

aka “named entity extraction”

Adapted from slide by William Cohen

What is Information Extraction

A family of techniques:

Information Extraction = segmentation + classification + association

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

Adapted from slide by William Cohen

What is Information Extraction

A family of techniques:

Information Extraction = segmentation + classification + association

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation CEO Bill Gates
Microsoft Gates
Microsoft
Bill Veghte Microsoft VP
Richard Stallman founder Free Software Foundation

Adapted from slide by William Cohen

What is Information Extraction

A family of techniques:

Information Extraction = segmentation + classification + association

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

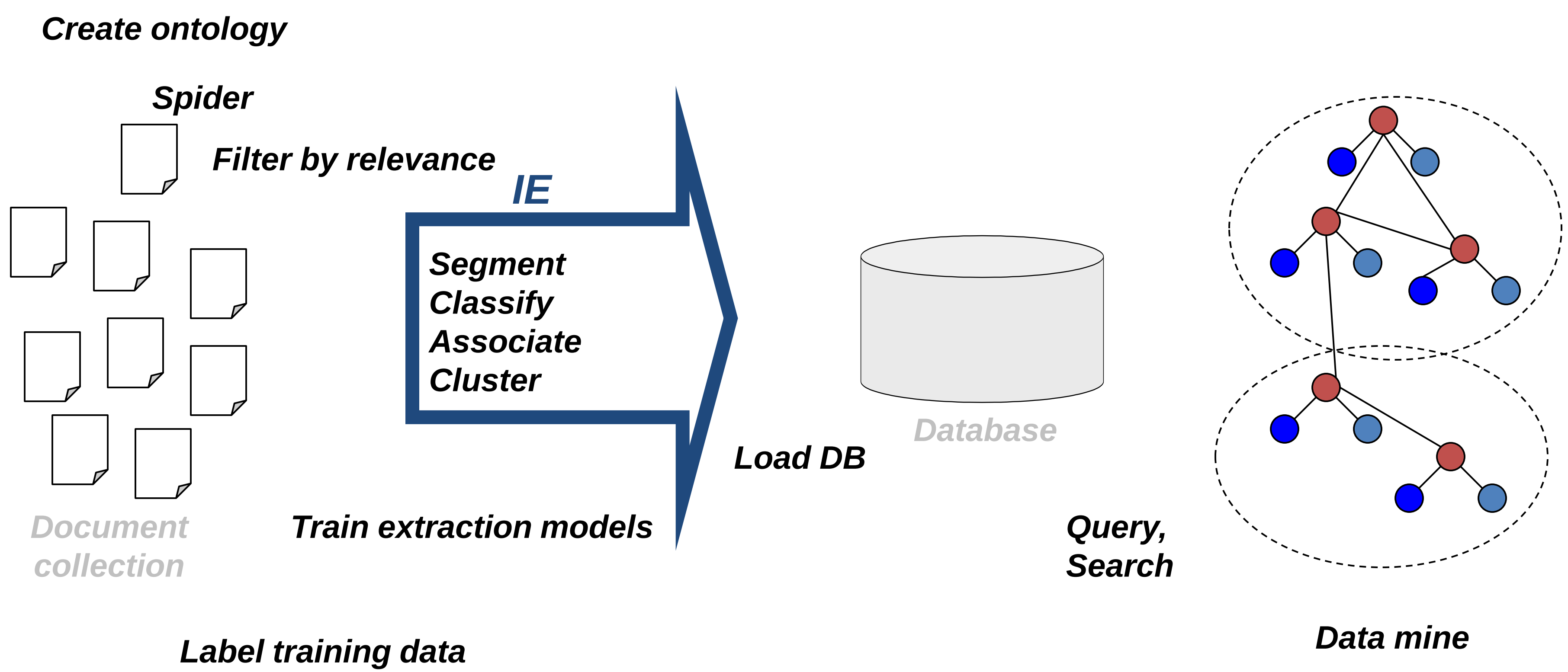
Microsoft Corporation CEO Bill Gates
Microsoft Gates

Microsoft

Bill Veghte Microsoft VP
Richard Stallman founder Free Software Foundation

Adapted from slide by William Cohen

IE in Context



Adapted from slide by William Cohen

Landscape of IE Tasks:

Degree of Formatting

innovate achieve lead

Text paragraphs without formatting

Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

Grammatical sentences and some formatting & links

Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

Frank Huybrechts - COO
Mr. Huybrechts has over 20 years of

- Press

Contact

- General information
- Directions maps

Non-grammatical snippets, rich formatting & links

Barto, Andrew G. (413) 545-2109 barto@cs.umass.edu CS276
Professor.
Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.

Berger, Emery D. (413) 577-4211 emery@cs.umass.edu CS344
Assistant Professor.

Brock, Oliver (413) 577-0334 oli@cs.umass.edu CS246
Assistant Professor.

Clarke, Lori A. (413) 545-1328 clarke@cs.umass.edu CS304
Professor.
Software verification, testing, and analysis; software architecture and design.

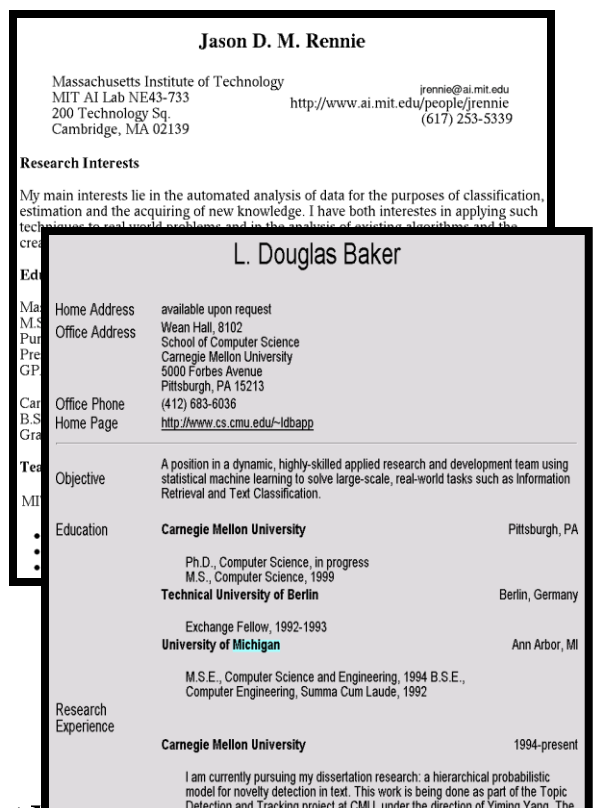
Cohen, Paul R. (413) 545-3638 cohen@cs.umass.edu CS278
Professor.
Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.

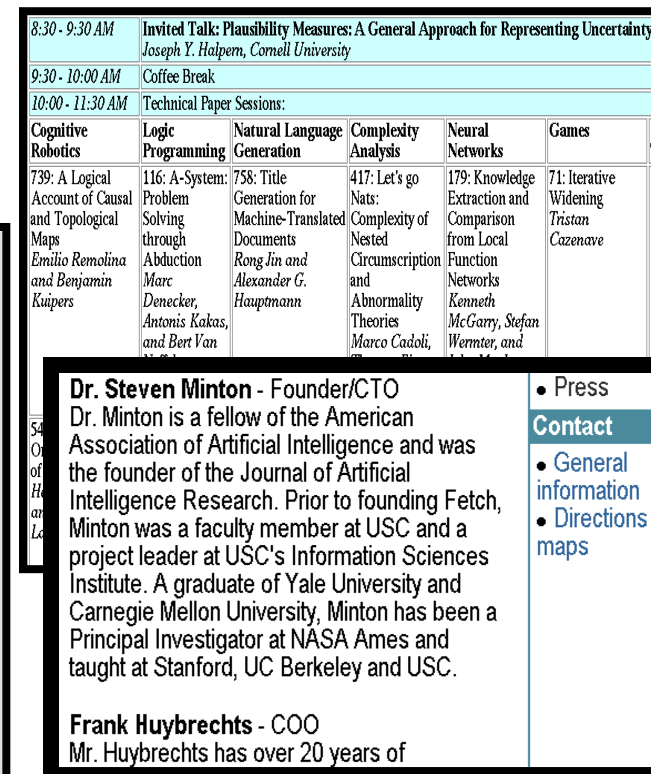
Tables

8:30 - 9:30 AM	**Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty** *Joseph Y. Halpern, Cornell University*					
9:30 - 10:00 AM	Coffee Break					
10:00 - 11:30 AM	Technical Paper Sessions:					
Cognitive Robotics	**Logic Programming**	**Natural Language Generation**	**Complexity Analysis**	**Neural Networks**	**Games**	
739: A Logical Account of Causal and Topological Maps *Emilio Remolina and Benjamin Kuipers*	116: A-System: Problem Solving through Abduction *Marc Denecker, Antonis Kakas, and Bert Van Nuffelen*	758: Title Generation for Machine-Translated Documents *Rong Jin and Alexander G. Hauptmann*	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories *Marco Cadoli, Thomas Eiter, and Georg Gottlob*	179: Knowledge Extraction and Comparison from Local Function Networks *Kenneth McGarry, Stefan Wermter, and John MacIntyre*	71: Iterative Widening *Tristan Cazenave*	
549: Online-Execution of ccGolog Plans *Henrik Grosskreutz*	131: A Comparative Study of Logic Programs with	246: Dealing with Dependencies between Content Planning and	470: A Perspective on Knowledge Compilation	258: Violation-Guided Learning for Constrained	353: Temporal Difference Learning Applied to a	

Landscape of IE Tasks:
Intended Breadth of Coverage

Web site specific	***Genre specific***	***Wide, non-specific***
Formatting	*Layout*	*Language*
Amazon.com Book Pages	Resumes	University Names





Adapted from slide by William Cohen

Information Extraction

- Information extraction (IE) systems
  - Find and understand limited 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:
    1. Organize information so that it is useful to people
    2. Put information in a semantically precise form that allows further inferences to be made by computer algorithms

Information Extraction

- 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-gene product interactions from medical research literature

Low level Information Extraction

- Is now available – and I think popular – in applications like Apple or Google mail, and web indexing

- Often seems to be based on regular expressions and name lists

Named Entity Recognition

- A very important sub-task: find and classify names in text, for example:
  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Named Entity Recognition

- A very important sub-task: find and classify names in text, for example:
  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:
  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Person
Date
Location
Organi-
zation

Evolution of NER

Traditional

- Rules
- Regular Expressions
- Gazetteers

Statistical

- Word-based models – PMI, log-likelihood.
- Sequence models – Conditional Random Fields

Neural

- Bi-LSTM
- Bi-LSTM+CRF
- Transformer based Models

Rule based NER

The textbook gives an example of an iterative approach that makes multiple passes over the text:

- Pass 1: Use high-precision rules
  to label (a small number of) unambiguous mentions
- Pass 2: Propagate the labels of the previously detected
  named entities to any mentions
  that are substrings (or acronyms?) of these entities
- Pass 3: Use application-specific name lists
  to identify further likely names (as features?)
- Pass 4: Now use a sequence labeling approach for NER,
  keeping the already labeled entities
  as high-precision anchors.

The basic ideas behind this approach (label propagation, using high-precision items as anchors) can be useful for other tasks as well.

NER Task

Task: Predict entities in a text

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

Standard evaluation is per entity, *not* per token

Variations and Ambiguity in NE

- Variation of NEs.
  - Manmohan Singh, Manmohan, Dr. Manmohan Singh
- Ambiguity of NE types:
  - 1945 (date vs. time)
  - Washington (location vs. person)
  - May (person vs. month)
  - Tata (person vs. organization)

More complex problems in NER

Issues of style, structure, domain, genre etc.

- Punctuation, spelling, spacing, formatting, ….all have an impact

Dept. of Computing and Information Science

Manchester Metropolitan University

Manchester

United Kingdom

> Tell me more about Leonardo

> Da Vinci

NER Approaches

Statistical models:

- Maximum Entropy Markov Models (MEMMs)
- Conditional Random Fields (CRFs)

Neural models:

- Recurrent networks (or transformers) that predict a label at each time step, possibly with a CRF output layer.

ML Sequence model Approach

Training

1. Collect a set of representative training documents
2. Label each token for its entity class or other (O)
3. Design feature extractors appropriate to the text and classes
4. Train a sequence classifier to predict the labels from the data

Testing

1. Receive a set of testing documents
2. Run sequence model inference to label each token
3. Appropriately output the recognized entities

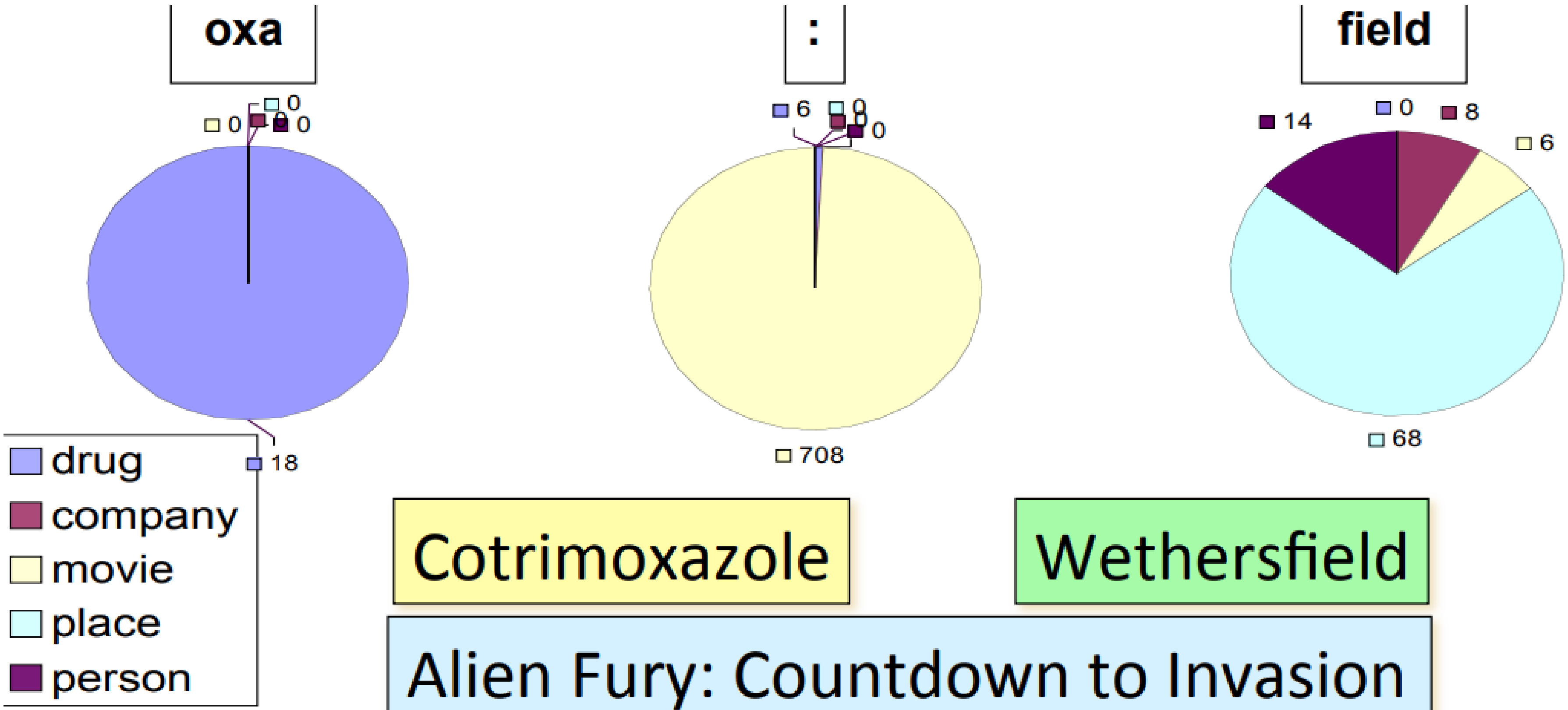
Encoding classes for sequence labeling

	IO encoding	IOB encoding
Fred	PER	B-PER
showed	O	O
Sue	PER	B-PER
Mengqiu	PER	B-PER
Huang	PER	I-PER
‘s	O	O
new	O	O
painting	O	O

Features of sequence labeling

- Words
  - Current word (essentially like a learned dictionary)
  - Previous/next word (context)
- Other kinds of inferred linguistic classification
  - Part-of-speech tags
- Label context
  - Previous (and perhaps next) label

Features: Word Substrings



Cotrimoxazole

Wethersfield

Alien Fury: Countdown to Invasion

Features: Word Shapes

- Word Shapes
  - Map words to simplified representation that encodes attributes such as length, capitalization, numerals, Greek letters, internal punctuation, etc.

Varicella-zoster	Xx-xxx
mRNA	xXXX
CPA1	XXXd

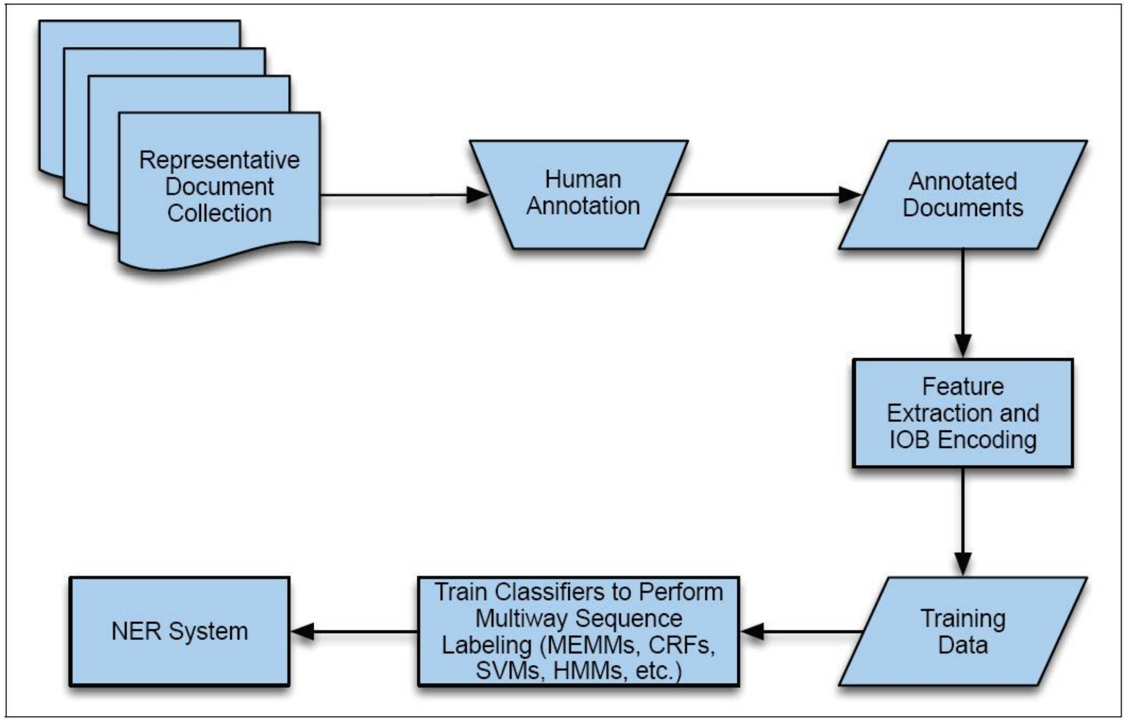
Named Entity Recognition

[ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad] .

IOB notation

Word	*POS*	*Chunk*	*EntityType*
U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

IE by statistical sequence labeling approach



```
Pierre Vinken , 61 years old , will join IBM 's board
as a nonexecutive director Nov. 29 .
```

```
[PERS Pierre Vinken] , 61 years old , will join
[ORG IBM] 's board as a nonexecutive director
[DATE Nov. 2] .
```

Task: identify all mentions of named entities (people, organizations, locations, dates)

We define many new tags:

- **B-PERS, B-DATE, ...:** beginning of a mention of a person/ date...
- **I-PERS, I-DATE, ...:** inside of a mention of a person/date...

```
[PERS Pierre Vinken] , 61 years old , will join
[ORG IBM] ‘s board as a nonexecutive director
[DATE Nov. 2] .
```

```
Pierre_B-PERS Vinken_I-PERS ,_O 61_O years_O old_O ,_O
will_O join_O IBM_B-ORG ‘s_O board_O as_O a_O
nonexecutive_O director_O Nov._B-DATE 29_I-DATE ._O
```

Biological NER

- There are a much wider range of entity types (semantic classes) in the biological domain

[$_{tissue}$ Plasma] [$_{GP}$ BNP] concentrations were higher in both the [$_{population}$ judo] and [$_{population}$ marathon groups] than in [$_{population}$ controls], and positively correlated with [$_{ANAT}$ LV] mass as well as with deceleration time.

Semantic class	Examples
Cell lines	*T98G, HeLa cell, Chinese hamster ovary cells, CHO cells*
Cell types	*primary T lymphocytes, natural killer cells, NK cells*
Chemicals	*citric acid, 1,2-diiodopentane, C*
Drugs	*cyclosporin A, CDDP*
Genes/proteins	*white, HSP60, protein kinase C, L23A*
Malignancies	*carcinoma, breast neoplasms*
Medical/clinical concepts	*amyotrophic lateral sclerosis*
Mouse strains	*LAFT, AKR*
Mutations	*C10T, Ala64 $\rightarrow$ Gly*
Populations	*judo group*

Source: Jurafsky & Martin "Speech and Language Processing"

Biological NER (cont.)

- NER in this domain is particularly difficult because of the various forms which the names can take:

  e.g. "insulin", "ether a go-go", "breast cancer associated 1"

  - Long names (thus multi-token boundary detection is needed)
  - Spelling/typographical variations
  - Abbreviations, symbols
  - (Of course) Ambiguity (common meaning or domain concepts)

- Extracted NEs are often mapped to **biomedical ontologies** (e.g. Gene Ontology, UMLS)

Source: Jurafsky & Martin "Speech and Language Processing"

Sequence Problems

- Many problems in NLP have data which is a sequence of characters, words, phrases, lines, or sentences …
- We can think of our task as one of labeling each item

VBG	NN	IN	DT	NN	IN	NN
Chasing	opportunity	in	an	age	of	upheaval

POS tagging

B	B	I	I	B	I	B	I	B	B
而	相	对	于	这	些	品	牌	的	价

Word segmentation

PERS	O	O	O	ORG	ORG
Murdoch	discusses	future	of	News	Corp.

Named entity recognition

Text segmentation

MEMM

- For a Conditional Markov Model (CMM) a.k.a. a Maximum Entropy Markov Model (MEMM), the classifier makes a single decision at a time, conditioned on evidence from observations and previous decisions
- A larger space of sequences is usually explored via search

Local Context

Decision Point

-3	-2	-1	0	+1
DT	NNP	VBD	???	???
The	Dow	fell	22.6	%

Features

W_0	22.6
W_{+1}	%
W_{-1}	fell
T_{-1}	VBD
T_{-1}-T_{-2}	NNP-VBD
hasDigit?	true
…	…

(Ratnaparkhi 1996; Toutanova et al. 2003, etc.)

MEMM: POS Tagging

- Scoring individual labeling decisions is no more complex than standard classification decisions
  - We have some assumed labels to use for prior positions
  - We use features of those and the observed data (which can include current, previous, and next words) to predict the current label

Local Context

Decision Point

-3	-2	-1	0	+1
DT	NNP	VBD	???	???
The	Dow	fell	22.6	%

Features

W_0	22.6
W_{+1}	%
W_{-1}	fell
T_{-1}	VBD
T_{-1}-T_{-2}	NNP-VBD
hasDigit?	true
…	…

(Ratnaparkhi 1996; Toutanova et al. 2003, etc.)

MEMM: POS Tagging

- POS tagging Features can include:
  - Current, previous, next words in isolation or together.
  - Previous one, two, three tags.
  - Word-internal features: word types, suffixes, dashes, etc.

Local Context

Decision Point

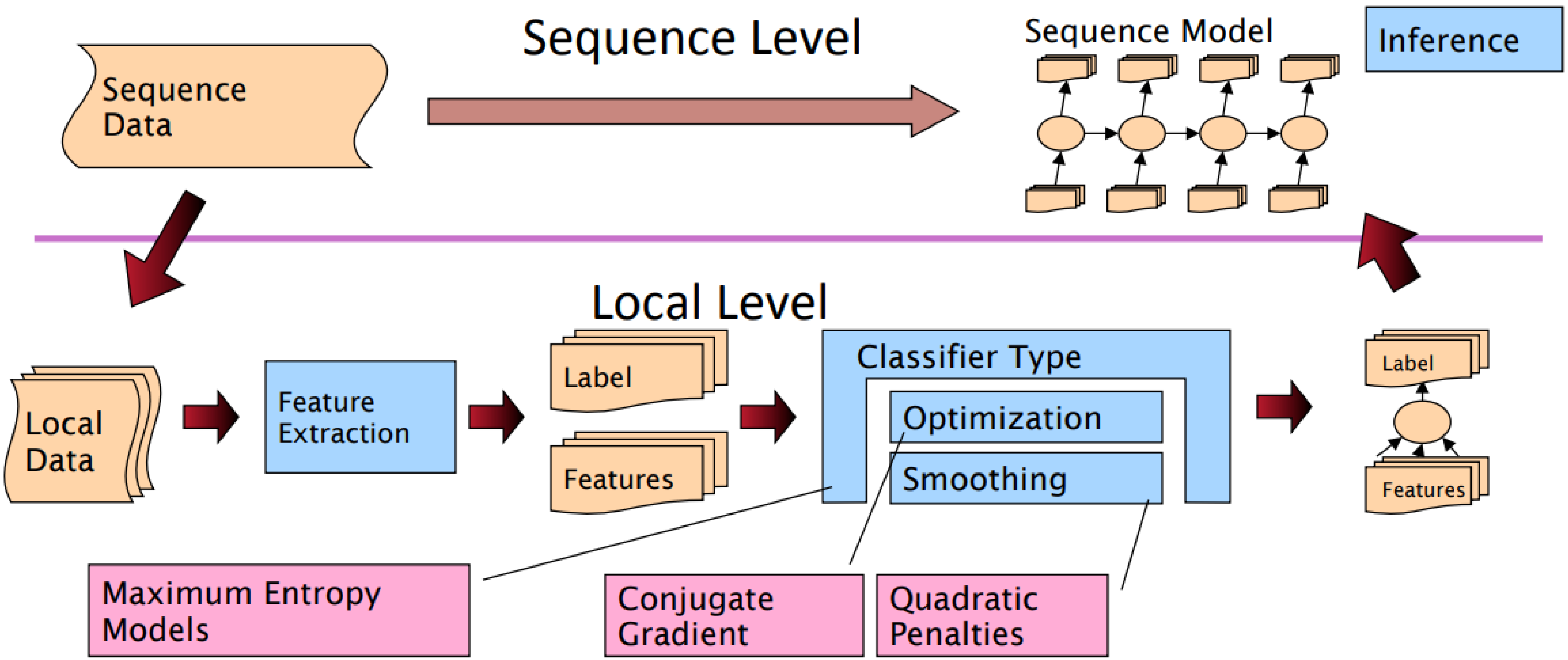
-3	-2	-1	0	+1
DT	NNP	VBD	???	???
The	Dow	fell	22.6	%

Features

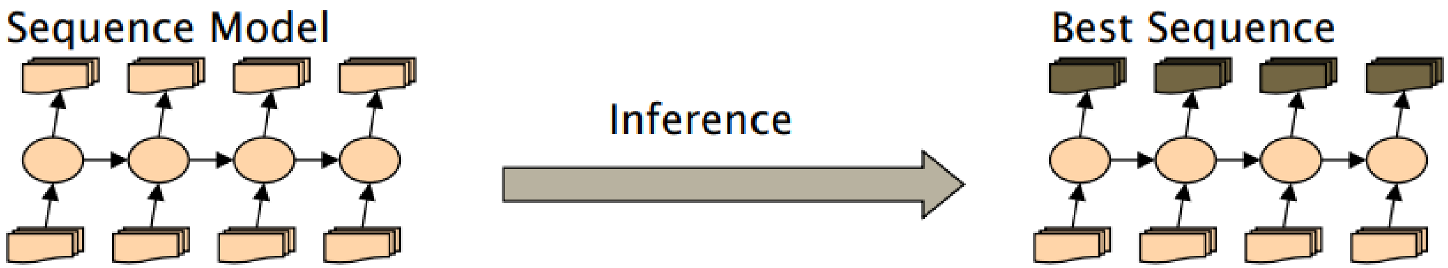
W_0	22.6
W_{+1}	%
W_{-1}	fell
T_{-1}	VBD
T_{-1}-T_{-2}	NNP-VBD
hasDigit?	true
…	…

(Ratnaparkhi 1996; Toutanova et al. 2003, etc.)

MEMM: Inference



MEMM: Greedy Inference



Greedy inference:

- We just start at the left, and use our classifier at each position to assign a label
- The classifier can depend on previous labeling decisions as well as observed data

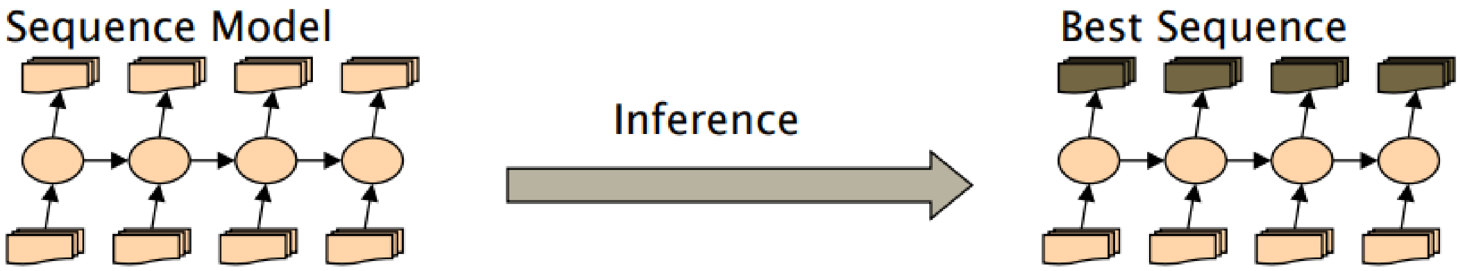
Advantages:

- Fast, no extra memory requirements
- Very easy to implement
- With rich features including observations to the right, it may perform quite well

Disadvantage:

- Greedy. We make commit errors we cannot recover from

MEMM: Beam Inference



Beam inference:

- At each position keep the top k complete sequences.
- Extend each sequence in each local way.
- The extensions compete for the k slots at the next position.

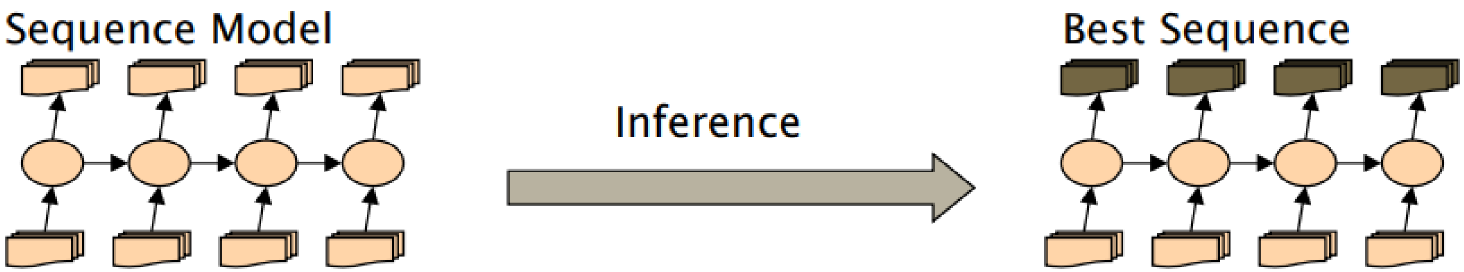
Advantages:

- Fast; beam sizes of 3-5 are almost as good as exact inference in many cases.
- Easy to implement (no dynamic programming required).

Disadvantage:

- Inexact: the globally best sequence can fall off the beam.

MEMM: Viterbi Inference



Viterbi inference:

- Dynamic programming or memoization.
- Requires small window of state influence (e.g., past two states are relevant).

Advantage:

- Exact: the global best sequence is returned.

Disadvantage:

- Harder to implement long-distance state-state interactions (but beam inference tends not to allow long-distance resurrection of sequences anyway).

Named Entity Types

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	**Turing** is a giant of computer science.
Organization	ORG	companies, sports teams	The **IPCC** warned about the cyclone.
Location	LOC	regions, mountains, seas	The **Mt. Sanitas** loop is in **Sunshine Canyon**.
Geo-Political Entity	GPE	countries, states, provinces	**Palo Alto** is raising the fees for parking.
Facility	FAC	bridges, buildings, airports	Consider the **Golden Gate Bridge**.
Vehicles	VEH	planes, trains, automobiles	It was a classic **Ford Falcon**.

Figure 18.1 A list of generic named entity types with the kinds of entities they refer to.

These types were developed for the news domain as part of NIST's Automatic Content Extraction (ACE) program.

Other domains (e.g. biomedical text) require different types (proteins, genes, diseases, etc.)

Feature based NER

identity of w_i, identity of neighboring words
embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
base-phrase syntactic chunk label of w_i and neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
w_i is all upper case
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
presence of hyphen

Figure 18.5 Typical features for a feature-based NER system.

Train a sequence labeling model (MEMM or CRF), using features such as the ones listed above for English

- Word Shape: replace all upper-case letters with one symbol (e.g. "X"), all lower-case letters with another symbol ("x"), all digits with another symbol ("d"), and leave punctuation marks as is ("L'Occitane → "X'Xxxxxxxx")
- Short Word Shape: remove adjacent letters that are identical in word shape "L'Occitane → "X'Xxxxxxxx" → "X'Xx")

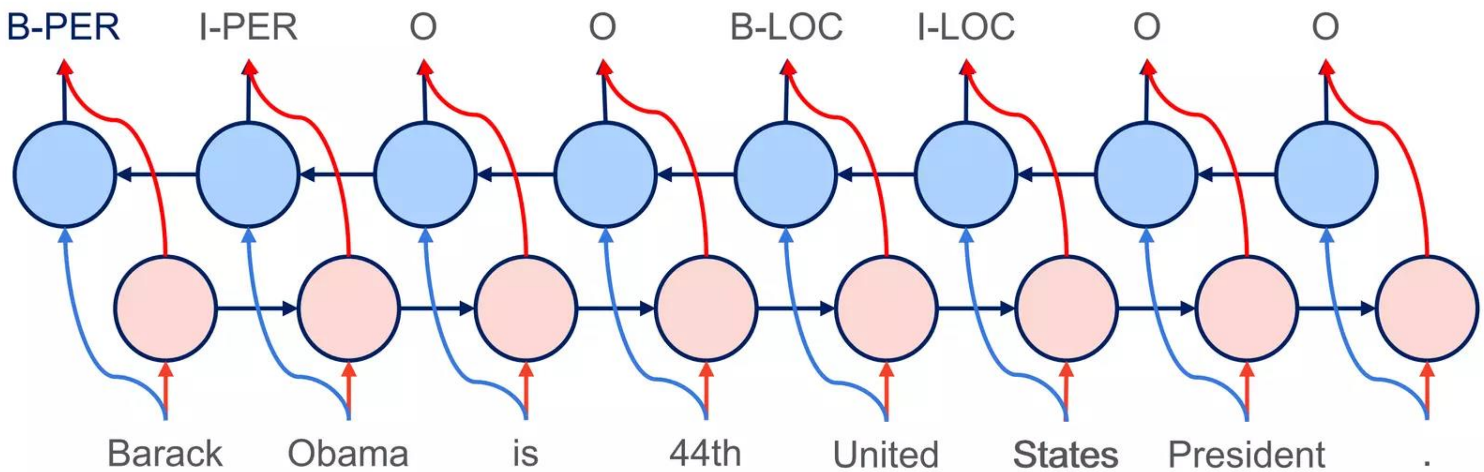
Input Format – BIO Tagging

- **BIO** – **B**egin **I**n **O**ut.
  - Barack/**B-PER** Obama/**I-PER** is/O 44th/O United/**B-LOC** States/**I-LOC** President/O ./O
- **BILOU – a tagging variant:**
  - **U** – **Unit** token (for single token entities)
  - **L** – **Last** token in sequence, ex. Barack/B-PER Obama/L-PER

CoNLL

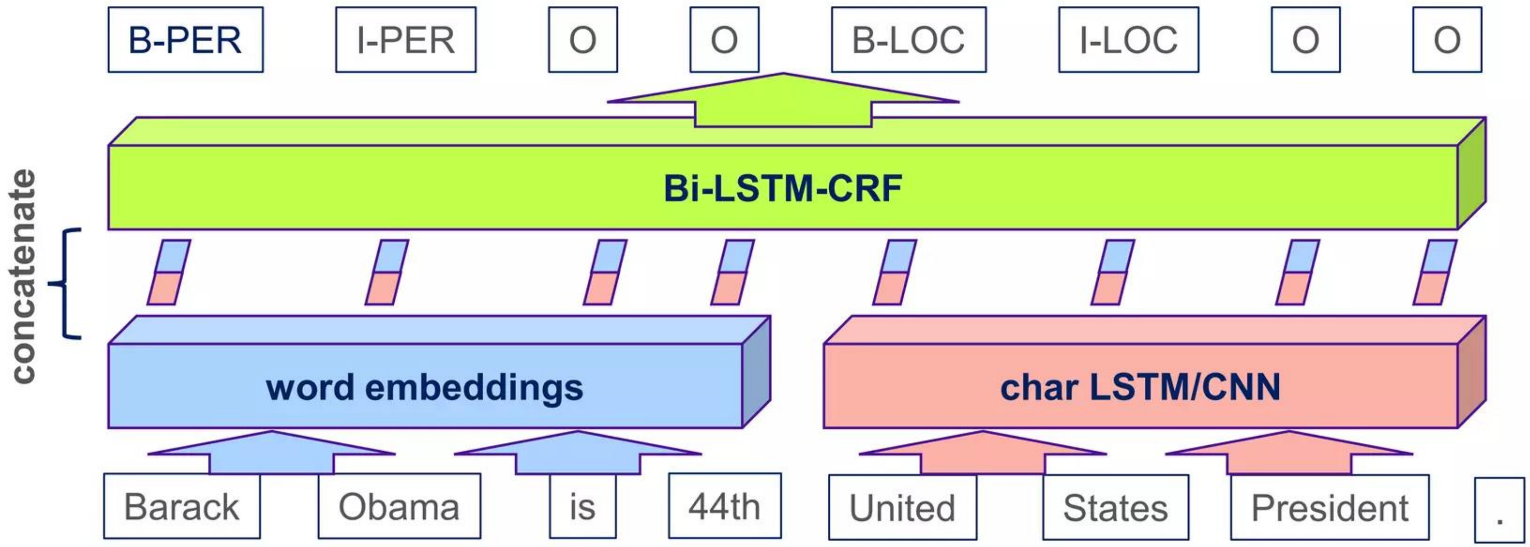
Barack	B-PER
Obama	I-PER
is	O
44th	O
United	B-LOC
States	I-LOC
President	O
.	O

Neural Model - BiLSTM



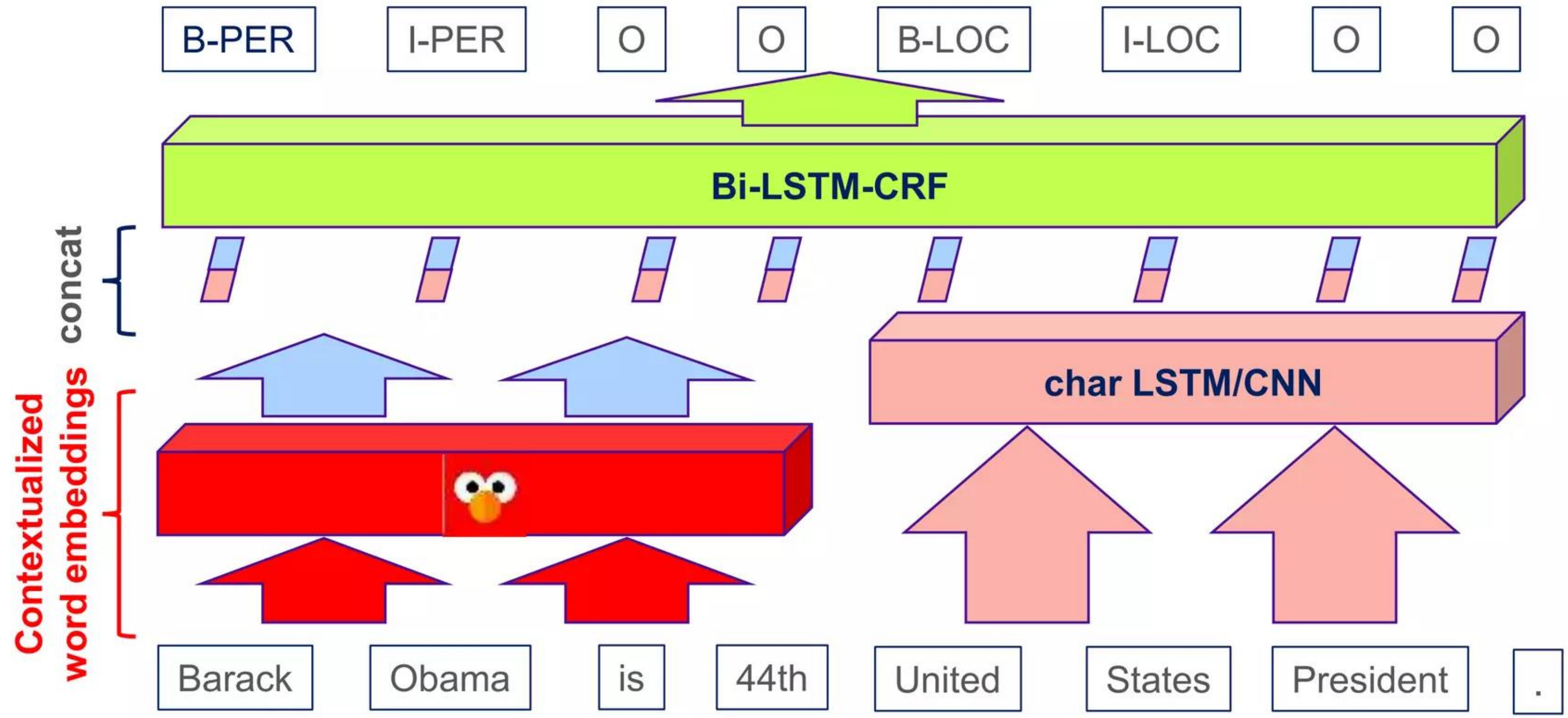
- Input is sequence of tokens, output is sequence of BIO tags.
- Weights trained end-to-end, no feature engineering needed.
- Bidirectional LSTM gets signal from neighboring words on both sides.

Neural Model – adding char embeddings

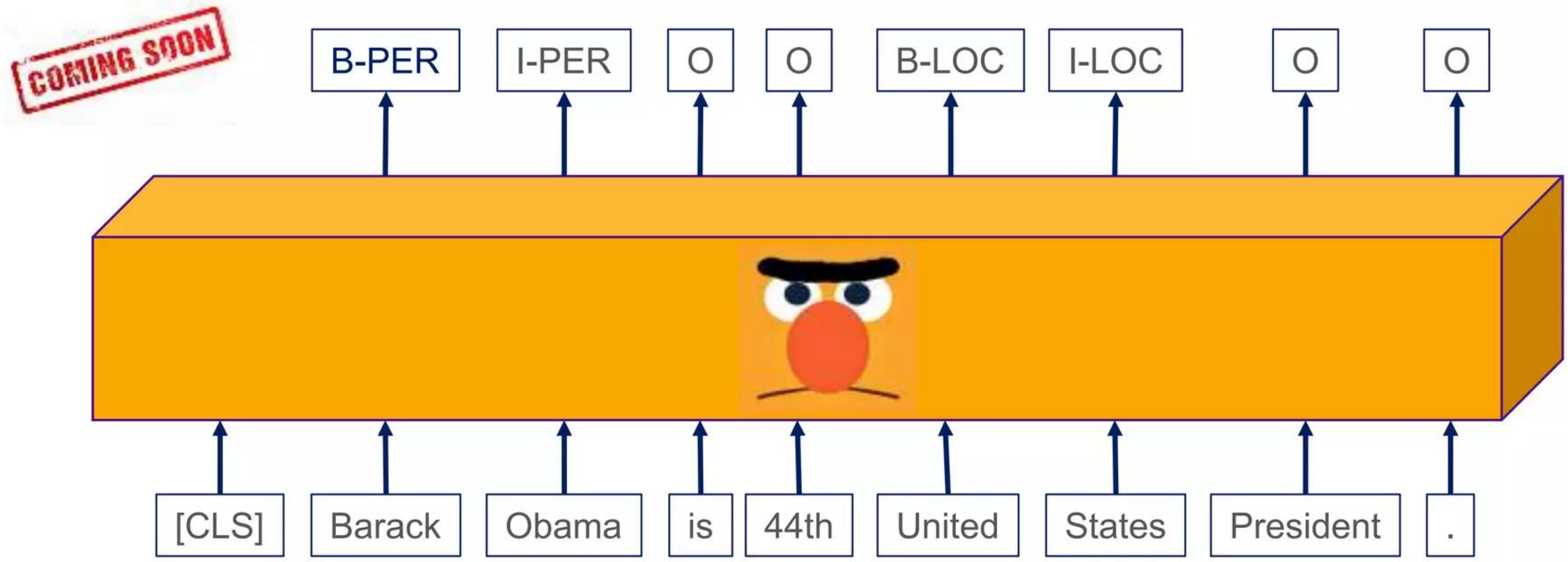


- Concatenate char embedding + word embedding and feed to Bi-LSTM-CRF.
- All weights learned end-to-end.
- Handles rare / unknown words; Exploits signal in prefix/suffix.

Neural Model – ELMo preprocessing

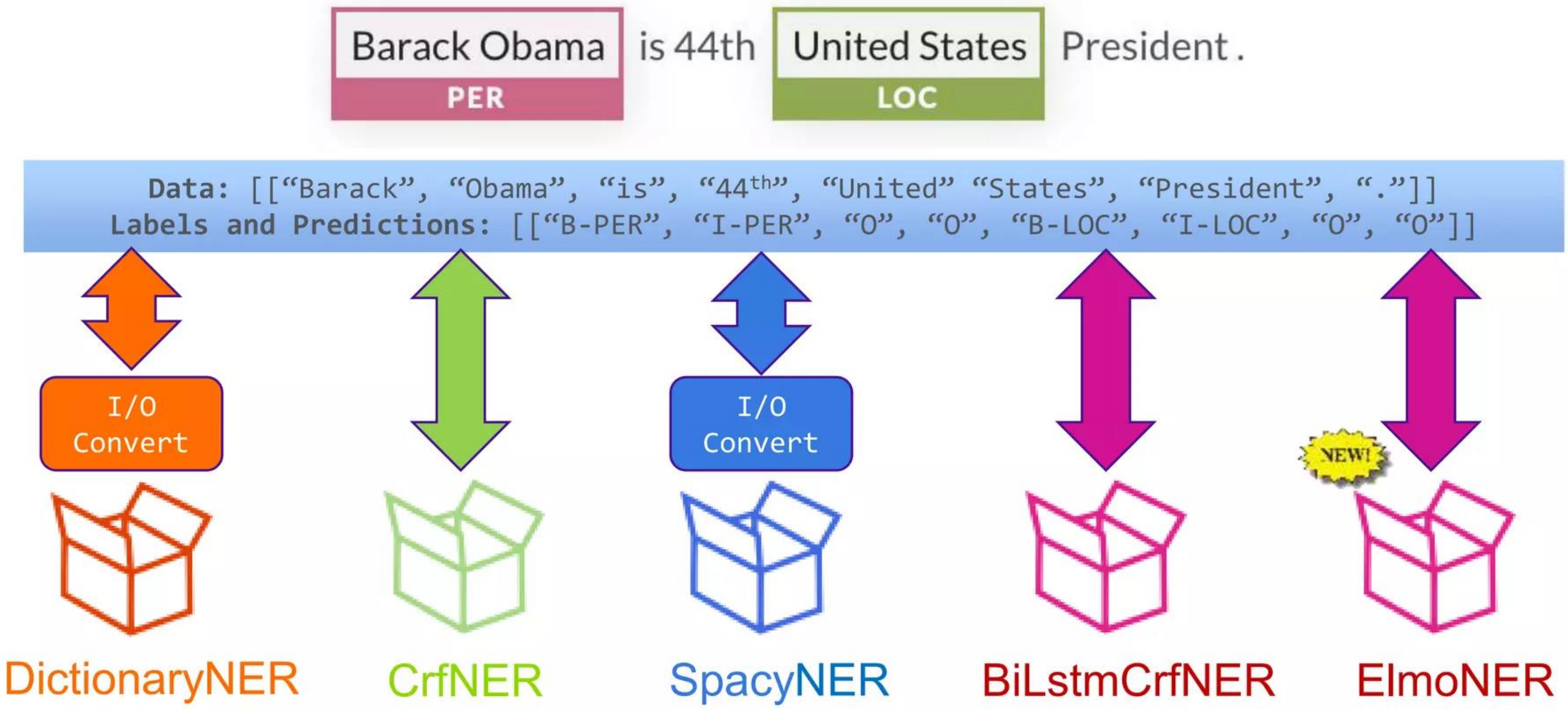


Neural Model – Transformer based



- BERT = Bidirectional Encoder Representation for Transformers.
- Source of embeddings similar to ELMo in standard BiLSTM + CRF models, OR
- Fine-tune LM backed NERs such as HuggingFace's BertForTokenClassification.

ELMo NER Model from Anago



Ensemble NER

```
# create and test an ensemble
dict_model = DictionaryNER()
dict_model.load("models/dict_model")
crf_model = CrfNER()
crf_model.load("models/crf_model")
spacy_model = SpacyNER()
spacy_model.load("models/spacy_model")
bilstm_model = BiLstmCrfNER()
bilstm_model.load("models/bilstm_model")
model = EnsembleNER()
model.fit(xtrain, ytrain,
    estimators=[
        (dict_model, {}),
        (crf_model, {}),
        (spacy_model, {}),
        (bilstm_model, {})
    ],
    is_pretrained=True)
ypred = model.predict(xtest)
print(classification_report(flatten_list(ytest, strip_prefix=True),
                            flatten_list(ypred, strip_prefix=True),
                            labels=entity_labels))
```

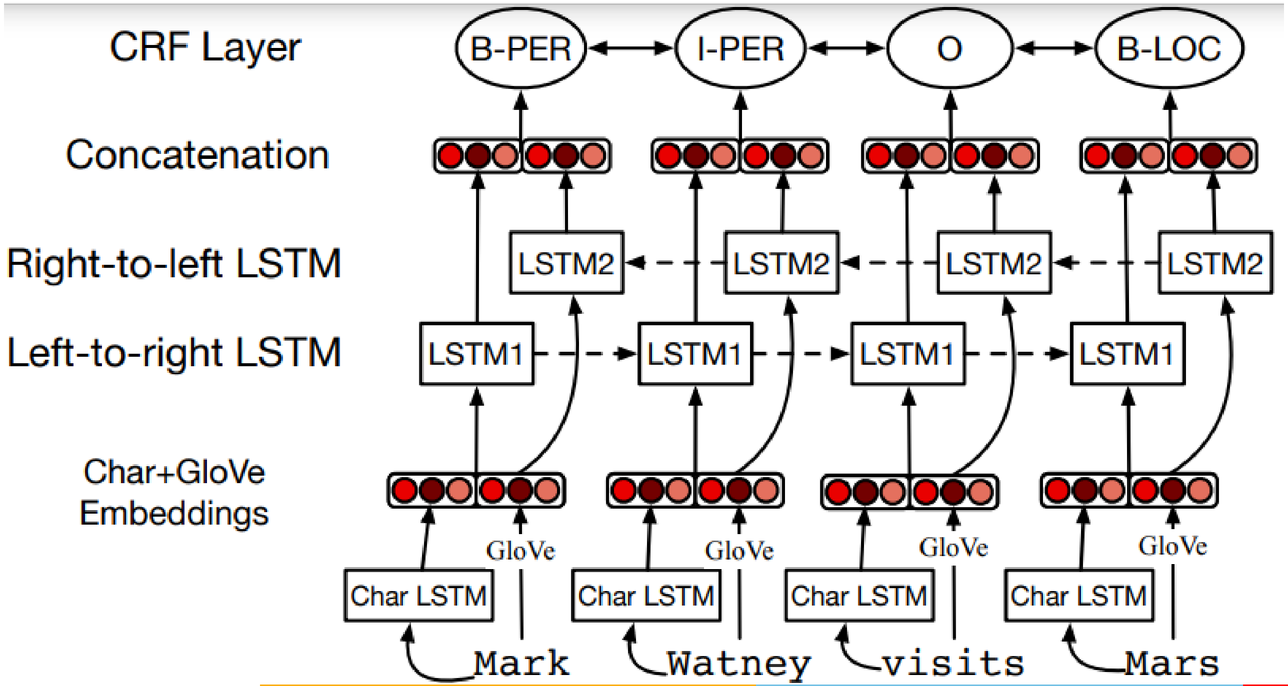
- Max Voting
- Improvements in this fork:
  - Unifies Max Voting and Weighted Max Voting NERs into single model.

ELSEVIER

Neural NER

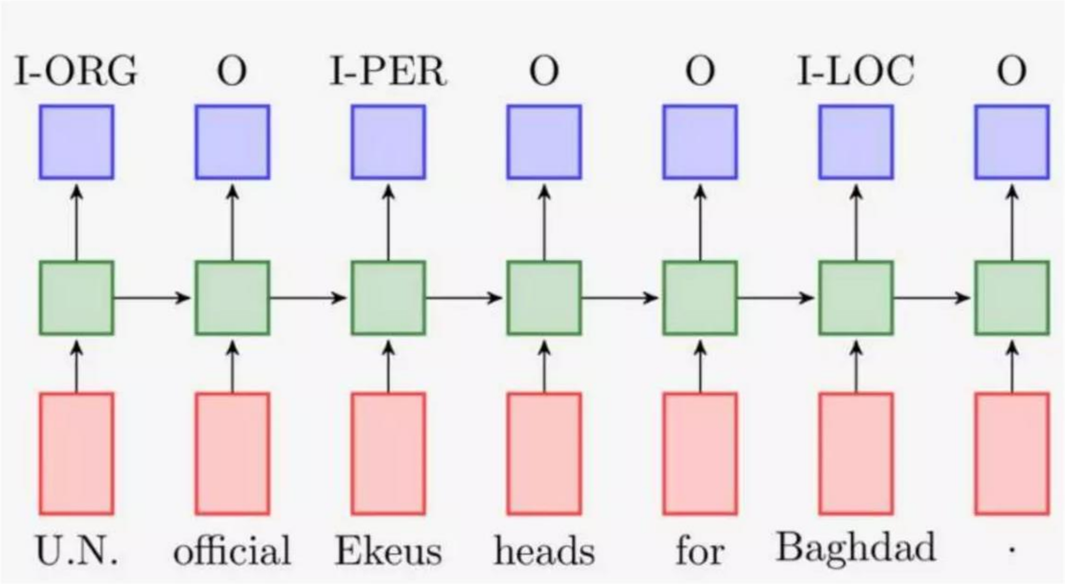
Sequence RNN (e.g. biLSTM or Transformer) with a CRF output layer.

Input: word embeddings, possibly concatenated with character embeddings and other features, e.g.:



Neural NER

- Feature extraction?
- Embeddings
- LSTM



Evaluating IE Accuracy

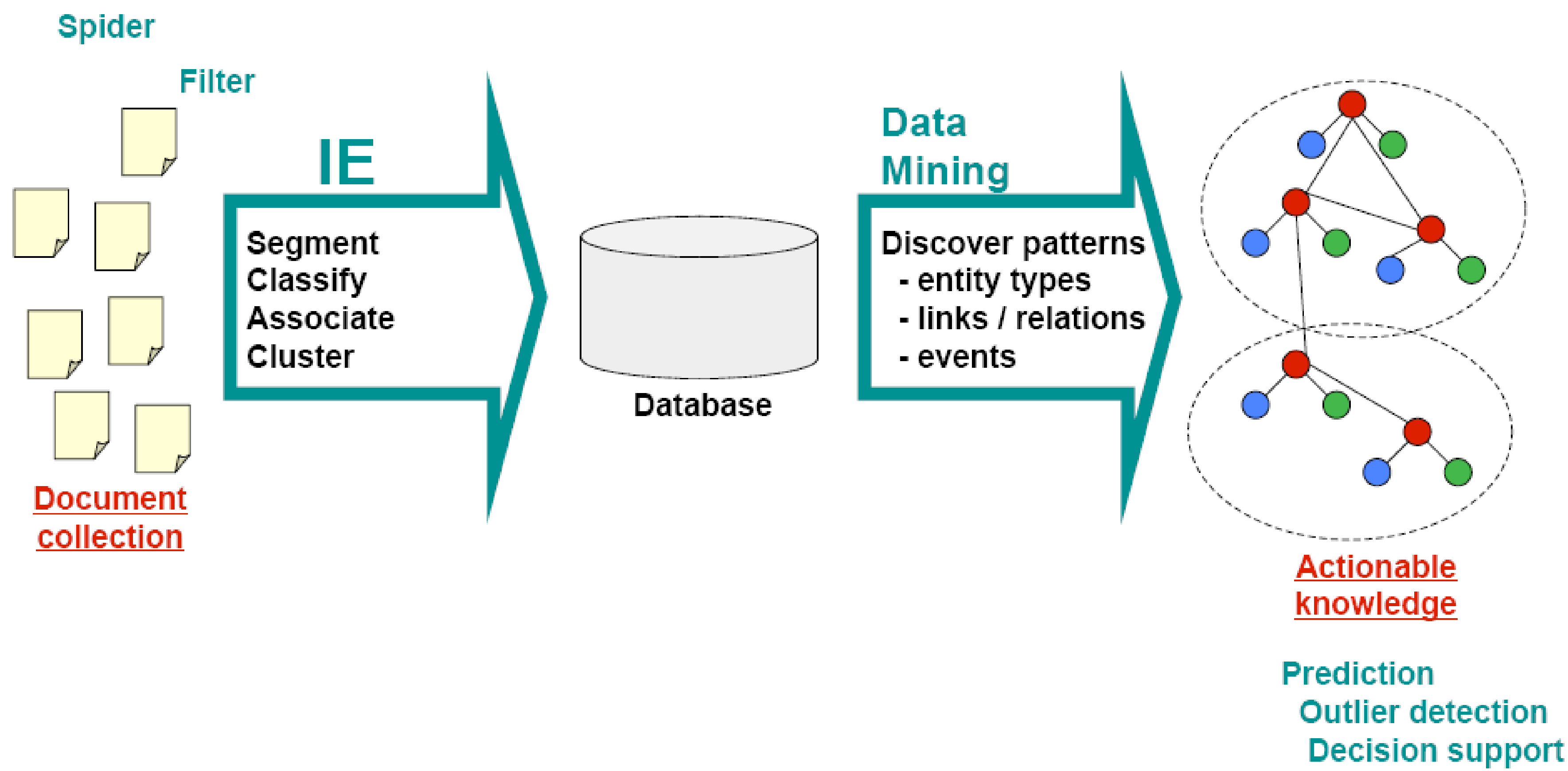
- Always evaluate performance on independent, manually-annotated test data not used during system development.
- Measure for each test document:
  - Total number of correct extractions in the solution template: *N*
  - Total number of slot/value pairs extracted by the system: *E*
  - Number of extracted slot/value pairs that are correct (i.e. in the solution template): *C*
- Compute average value of metrics adapted from IR:
  - Recall = *C*/*N*
  - Precision = *C*/*E*
  - F-Measure = Harmonic mean of recall and precision

Slide by Chris Manning, based on slides by others

Precision, Recall, F1 for NER

- Recall and precision are straightforward for tasks like IR and text categorization, where there is only one grain size (documents)
- The measure behaves a bit funnily for IE/NER when there are *boundary errors* (which are *common*):
  - First Bank of Chicago announced earnings …
- This counts as both a fp and a fn
- Selecting *nothing* would have been better
- Some other metrics (e.g., MUC scorer) give partial credit (according to complex rules)

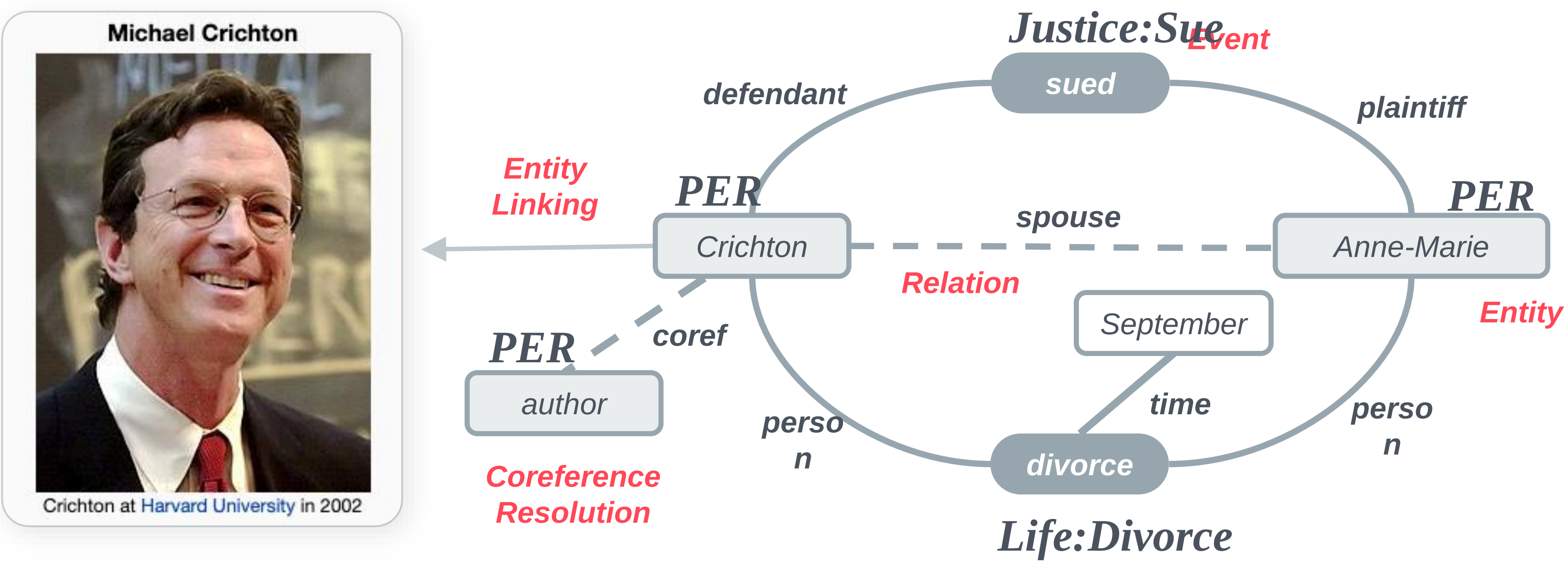
From Text to Actionable Knowledge



Source: Andrew McCallum, UMass Amherst

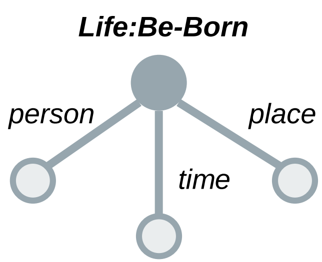
A Quick Overview of Information Extraction

“ ***Anne-Marie*** *sued* ***Crichton****, best known as the* ***author*** *of Jurassic Park, for divorce in September.*



A Brief Introduction to Information Extraction Subtasks

- **Entity Extraction** aims to identify entity mentions in text and classify them into pre-defined entity types.
- **Relation Extraction** is the task of assigning a relation type to an ordered pair of entity mentions.
- **Event Extraction** entails identifying and classifying event triggers and their arguments
  - Event triggers: the words or phrase that most clearly express event occurrences
  - Arguments: the words or phrases for participants in those events



- **Entity Coreference Resolution** is the task of resolving all entity mentions that refer to the same entity.

Relation Extraction

- Up until now we have focused on early stages of the Information Extraction pipeline
  - We have emphasized named entity tagging
- Now we will discuss extracting facts about these entities
  - This can include IS-A facts (similar to named entity types), but also more complicated relations

Extracting relations from text

- Company report: "International Business Machines Corporation (IBM or the company) was incorporated in the State of New York on June 16, 1911, as the Computing-Tabulating-Recording Co. (C-T-R)…"
- Extracted Complex Relation:

  Company-Founding

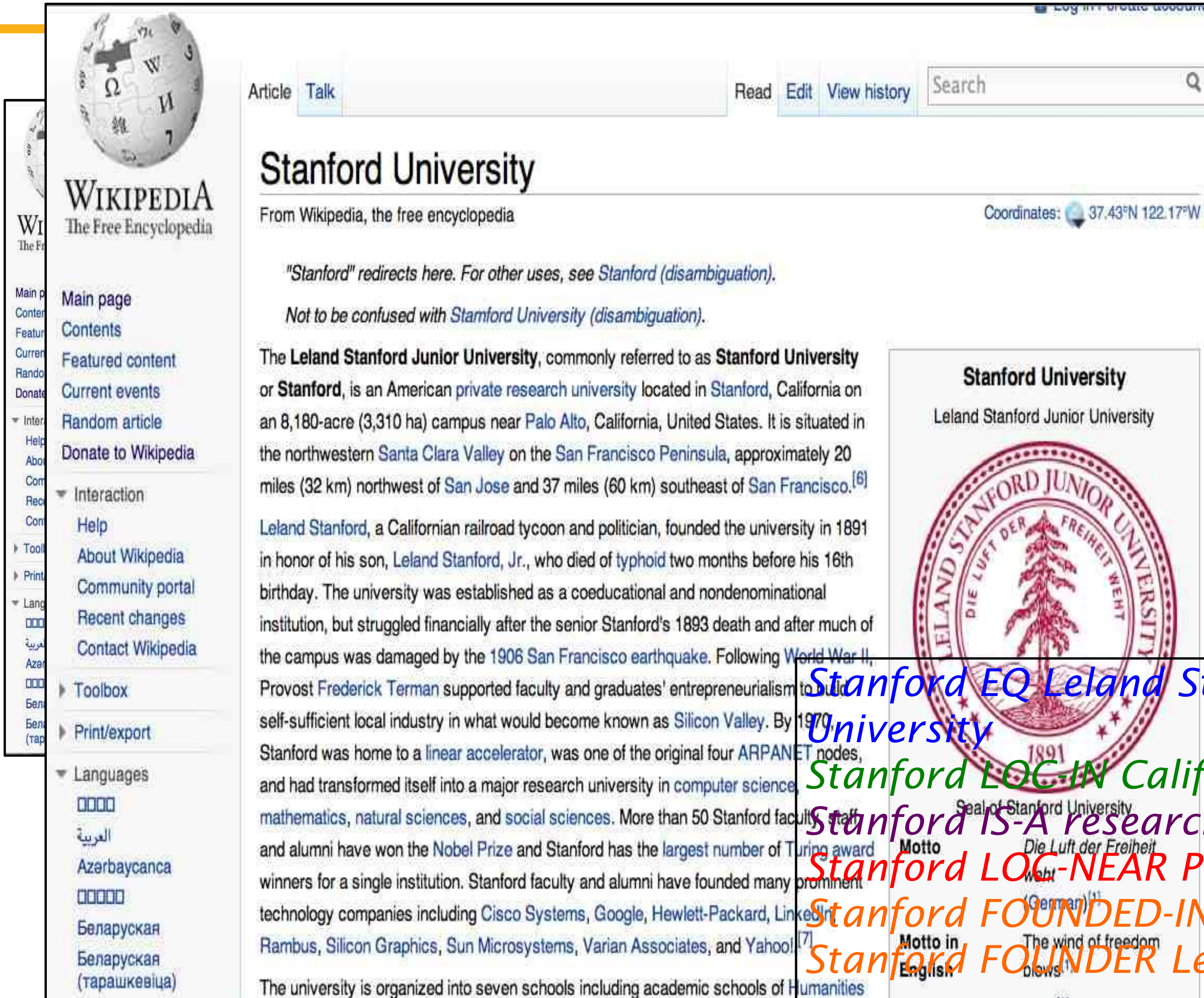
Company	IBM
Location	New York
Date	June 16, 1911
Original-Name	Computing-Tabulating-Recording Co.

- But we will focus on the simpler task of extracting relation **triples**

  Founding-year(IBM, 1911)

  Founding-location(IBM, New York)

Extracting Relation Triples from Text


Article Talk Read Edit View history Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox
Print/export
Languages
العربية
Azərbaycanca
Беларуская
Беларуская (тарашкевіца)

Stanford University

From Wikipedia, the free encyclopedia

Coordinates: 37.43°N 122.17°W

"Stanford" redirects here. For other uses, see Stanford (disambiguation).

Not to be confused with Stamford University (disambiguation).

The **Leland Stanford Junior University**, commonly referred to as **Stanford University** or **Stanford**, is an American private research university located in Stanford, California on an 8,180-acre (3,310 ha) campus near Palo Alto, California, United States. It is situated in the northwestern Santa Clara Valley on the San Francisco Peninsula, approximately 20 miles (32 km) northwest of San Jose and 37 miles (60 km) southeast of San Francisco.[6]

Leland Stanford, a Californian railroad tycoon and politician, founded the university in 1891 in honor of his son, Leland Stanford, Jr., who died of typhoid two months before his 16th birthday. The university was established as a coeducational and nondenominational institution, but struggled financially after the senior Stanford's 1893 death and after much of the campus was damaged by the 1906 San Francisco earthquake. Following World War II, Provost Frederick Terman supported faculty and graduates' entrepreneurialism to build self-sufficient local industry in what would become known as Silicon Valley. By 1970, Stanford was home to a linear accelerator, was one of the original four ARPANET nodes, and had transformed itself into a major research university in computer science, mathematics, natural sciences, and social sciences. More than 50 Stanford faculty, staff and alumni have won the Nobel Prize and Stanford has the largest number of Turing award winners for a single institution. Stanford faculty and alumni have founded many prominent technology companies including Cisco Systems, Google, Hewlett-Packard, LinkedIn, Rambus, Silicon Graphics, Sun Microsystems, Varian Associates, and Yahoo![7]

The university is organized into seven schools including academic schools of Humanities

Stanford University
Leland Stanford Junior University
Seal of Stanford University
Motto: Die Luft der Freiheit weht (German)[1]
Motto in English: The wind of freedom blows[1]

...rd Junior
...nly referred to as
... or Stanford, is an
...esearch university
..., California … near
...ia… Leland
...the university in

Stanford EQ Leland Stanford Junior University
Stanford LOC-IN California
Stanford IS-A research university
Stanford LOC-NEAR Palo Alto
Stanford FOUNDED-IN 1891
Stanford FOUNDER Leland Stanford

Slide from Dan Jurafsky

Why Relation Extraction?

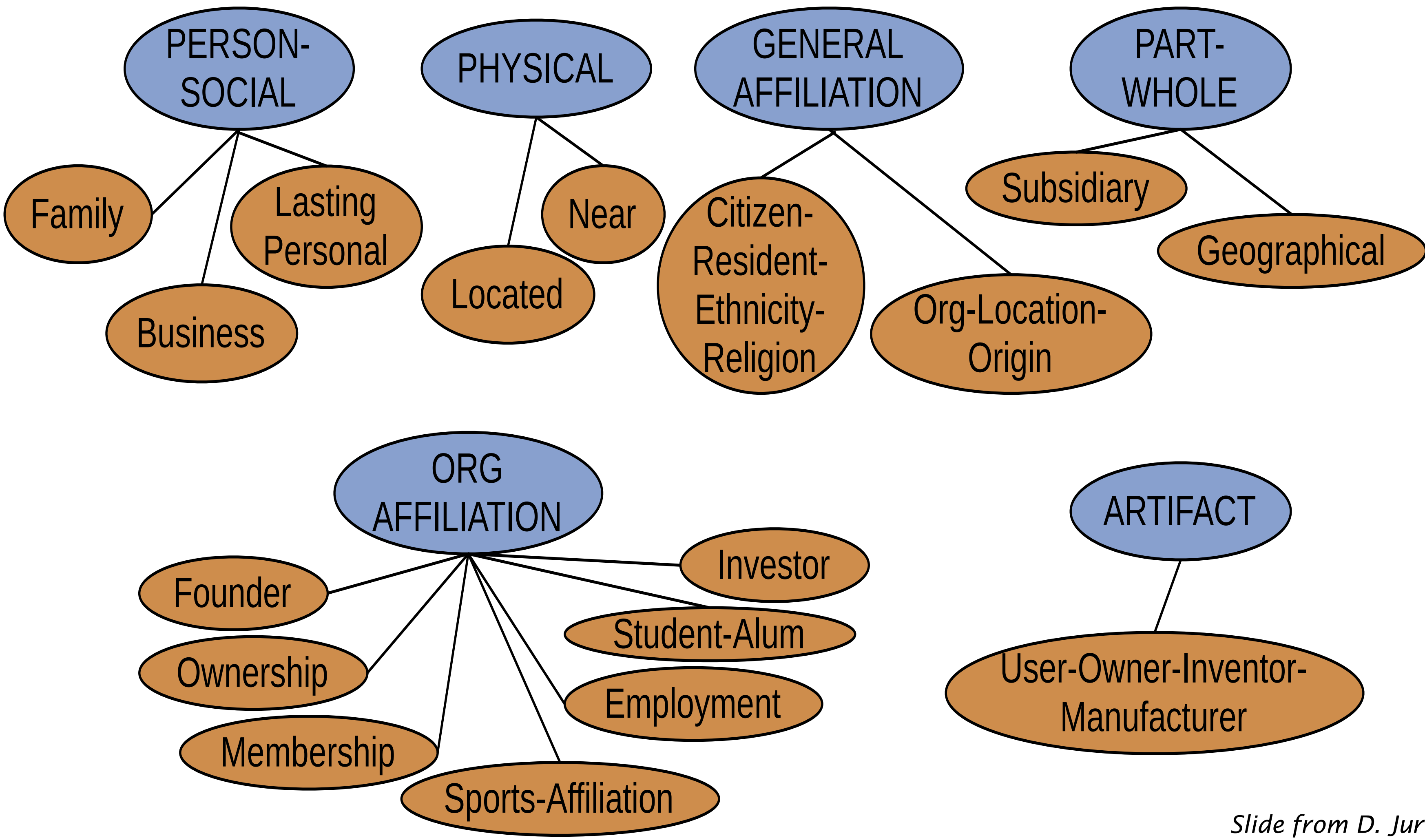
- Create new structured knowledge bases, useful for any app
- Augment current knowledge bases
- Adding words to WordNet thesaurus, facts to FreeBase or DBPedia
- Support ques on answering

The granddaughter of which actor starred in the movie “E.T.”? (acted-in ?x “E.T.”)(is-a ?y actor)(granddaughter-of ?x ?y)

- But which relations should we extract?

Automated Content Extraction (ACE)

17 relations from 2008 "Relation Extraction Task"



Slide from D. Jurafsky

Automated Content Extraction (ACE)

- Physical-Located PER-GPE

  He was in Tennessee

- Part-Whole-Subsidiary ORG-ORG

  XYZ, the parent company of ABC

- Person-Social-Family PER-PER

  John's wife Yoko

- Org-AFF-Founder PER-ORG

  Steve Jobs, co-founder of Apple…

Slide from Dr. Jurafsky

UMLS: Unified Medical Language System

- 134 entity types, 54 relations

Injury	*disrupts*	*Physiological Function*
Bodily Location	*location-of*	*Biologic Function*
Anatomical Structure	*part-of*	*Organism*
Pharmacologic Substance	*causes*	*Pathological Function*
Pharmacologic Substance	*treats*	*Pathologic Function*

Slide from D. Jurafsky

Extracting UMLS relations from a sentence

```
Doppler echocardiography can be used
to diagnose left anterior descending
artery stenosis in patients with type 2
diabetes
```

↓

Echocardiography, Doppler DIAGNOSES Acquired stenosis

Slide from Dr. Jurafsky

Databases of Wikipedia Relations

Wikipedia Infobox

```
{{Infobox university
|image_name= Stanford University seal.svg
|image_size= 210px
|caption = Seal of Stanford University
|name =Stanford University
|native_name =Leland Stanford Junior University
|motto = {{lang|de|"Die Luft der Freiheit weht"}}<br /> ([[German language|German]])<ref
name="casper">{{cite speech|title=Die Luft der Freiheit weht—On and Off|author=Gerhard
Casper|first=Gerhard|last=Casper|authorlink=Gerhard Casper|date=1995-10-
05|url=http://www.stanford.edu/dept/pres-provost/president/speeches/951005dieluft.html}}</ref>
|mottoeng = The wind of freedom blows<ref name="casper" />
|established = 1891<ref>{{cite web |
url=http://www.stanford.edu/home/stanford/history/begin.html | title=Stanford University History |
publisher = Stanford University | accessdate = 2007-04-26}}</ref>
|type = [[private university|Private]]
|calendar= Quarter
|president = [[John L. Hennessy]]
|provost = [[John Etchemendy]]
|city = [[Stanford, California|Stanford]]
|state = California
|country = U.S.
```

Relations extracted from Infobox

Stanford state California

Stanford motto "Die Luft der Frei

…

Relation databases that draw from Wikipedia

- Resource Description Framework (RDF) triples

  subject predicate object

  Golden Gate Park location San Francisco

  dbpedia:Golden_Gate_Park dbpedia-owl:location dbpedia:San_Francisco

- DBPedia: 1 billion RDF triples, 385 from English Wikipedia
- Frequent Freebase relations:

  people/person/nationality, location/location/contains

  people/person/profession, people/person/place-of-birth

  biology/organism_higher_classification film/film/genre

Slide from Dan Jurafsky

Ontological relations

Examples from the WordNet Thesaurus

- IS-A (hypernym): subsumption between classes
  - `Giraffe` IS-A `ruminant` IS-A `ungulate` IS-A `mammal` IS-A `vertebrate` IS-A `animal…`
- Instance-of: relation between individual and class
  - `San Francisco` instance-of `city`

Slide from Dr. Jurafsky

Patterns for Relation Extraction

- Hand-written rules for relation extraction were used in MUC (such as the Fastus system)
- Recently there has been a renewed wide interest in learning rules for relation extraction focused on precision
  - The presumption is that interesting information occurs many times on the web, with different contexts
    - e.g., how many times does "Barack Obama is the 44th President of the United States" occur on the web?
  - Focusing on high precision is reasonable because the high redundancy will allow us to deal with recall

How to build relation extractors

1. Hand-written patterns
2. Supervised machine learning
3. Semi-supervised and unsupervised
   - Bootstrapping (using seeds)
   - Distant supervision
   - Unsupervised learning from the web
4. Deep Learning

Slide from Dr. Jurafsky

Rules for extracting IS-A relation

Early intuition from **Hearst (1992)**

- "Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use"

- What does *Gelidium* mean?
- How do you know?`

Rules for extracting IS-A relation

Early intuition from **Hearst (1992)**

- “Agar is a substance prepared from a mixture of **red algae, such as Gelidium,** for laboratory or industrial use”

- What does *Gelidium* mean?
- How do you know?`

Hearst’s Patterns for extracting IS-A relations

(Hearst, 1992): Automatic Acquisition of Hyponyms

“Y such as X ((, X)* (, and|or) X)”
“such Y as X”
“X or other Y”
“X and other Y”
“Y including X”
“Y, especially X”

Hearst's Patterns for extracting IS-A relations

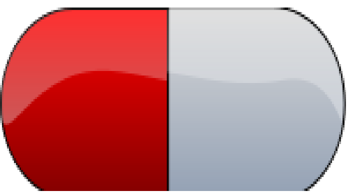
Hearst pattern	Example occurrences
X and other Y	...temples, treasuries, and other important civic buildings.
X or other Y	Bruises, wounds, broken bones or other injuries...
Y such as X	The bow lute, such as the Bambara ndang...
Such Y as X	...such authors as Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, including Canada and England...
Y , especially X	European countries, especially France, England, and Spain...

Extracting Richer Relations Using Rules

- Intuition: relations often hold between specific entities
  - located-in (ORGANIZATION, LOCATION)
  - founded (PERSON, ORGANIZATION)
  - cures (DRUG, DISEASE)
- Start with Named Entity tags to help extract relation!

Which relations hold between 2 entities?

Named Entities aren't quite enough.



Cure?

Prevent?

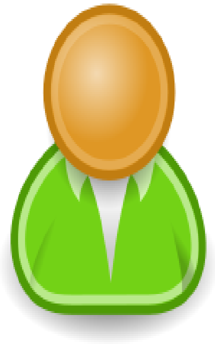
Cause?

Drug

Disease

Slide from Dan Jurafsky

What relations hold between 2 entities?



PERSON

Founder?

Investor?

Member?

Employee?

President?

ORGANIZATION

Slide from D. Jurafsky

Extracting Richer Relations Using Rules and Named Entities

Who holds what office in what organization?

PERSON, POSITION of ORG

- George Marshall, Secretary of State of the United States

PERSON (named|appointed|chose|*etc.*) PERSON Prep? POSITION

- Truman appointed Marshall Secretary of State

PERSON [be]? (named|appointed|*etc.*) Prep? ORG POSITION

- George Marshall was named US Secretary of State

Idea: define some extraction patterns

X is the founder of Y	→	48-year-old Elon Musk is the founder of SpaceX and a co-founder of Tesla Motors.
X, who founded Y	→	Elon Musk, who founded SpaceX in 2002, has said the company is focused on ...
Y was founded by X	→	SpaceX was founded by Elon Musk to make life multi-planetary. "You want to ...

Problem: most occurrences do not fit simple patterns

You may also be thinking of Elon Musk (founder of SpaceX), who started PayPal.

Elon Musk, co-founder of PayPal, went on to establish SpaceX, one of the most ...

If Space Exploration (SpaceX), founded by Paypal pioneer Elon Musk succeeds, ...

Hand-built patterns for relations

- Plus:
  - Human patterns tend to be high-precision
  - Can be tailored to specific domains
- Minus
  - Human patterns are often low-recall
  - A lot of work to think of all possible patterns!
  - Don’t want to have to do this for every relation!
  - We’d like better accuracy

Slide from Dr. Jurafsky

Supervised Methods

- For named entity tagging, statistical taggers are the state of the art
- However, for relation extraction, this is not necessarily true
  - Still many hand-crafted rule-based systems out there that work well
  - But hand-crafting such systems takes a lot of work, so classification approaches are very interesting (and they are improving with time)
- Formulate relation extraction as a supervised classification problem

Supervised machine learning for relations

- Choose a set of relations we'd like to extract
- Choose a set of relevant named entities
- Find and label data
  - Choose a representative corpus
  - Label the named entities in the corpus
  - Hand-label the relations between these entities
  - Break into training, development, and test
- Train a classifier on the training set

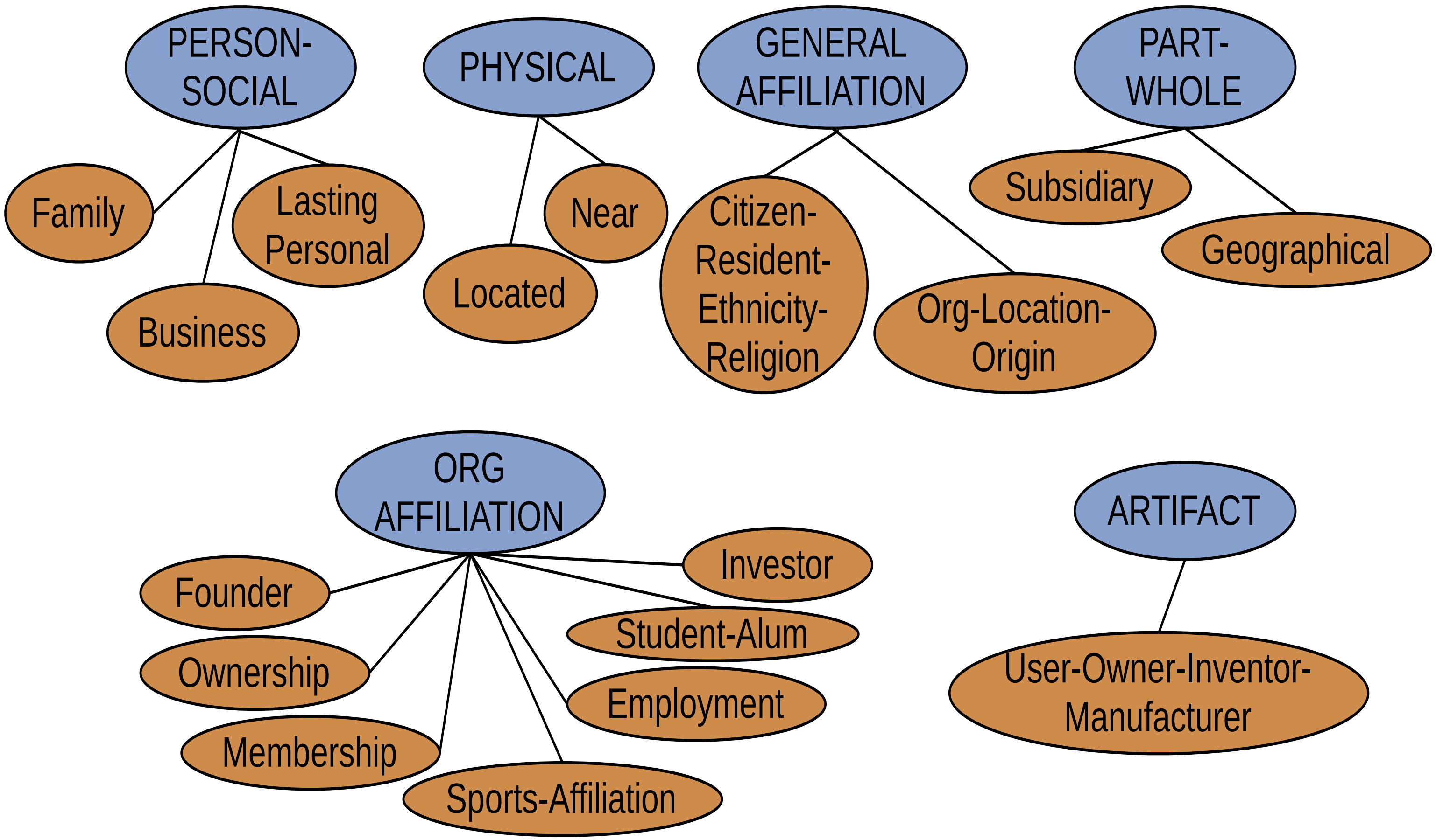
How to do classification in supervised relation extraction

1. Find all pairs of named entities (usually in same sentence)
2. Decide if 2 entities are related
3. If yes, classify the relation

- Why the extra step?
  - Faster classification training by eliminating most pairs
  - Can use distinct feature-sets appropriate for each task.

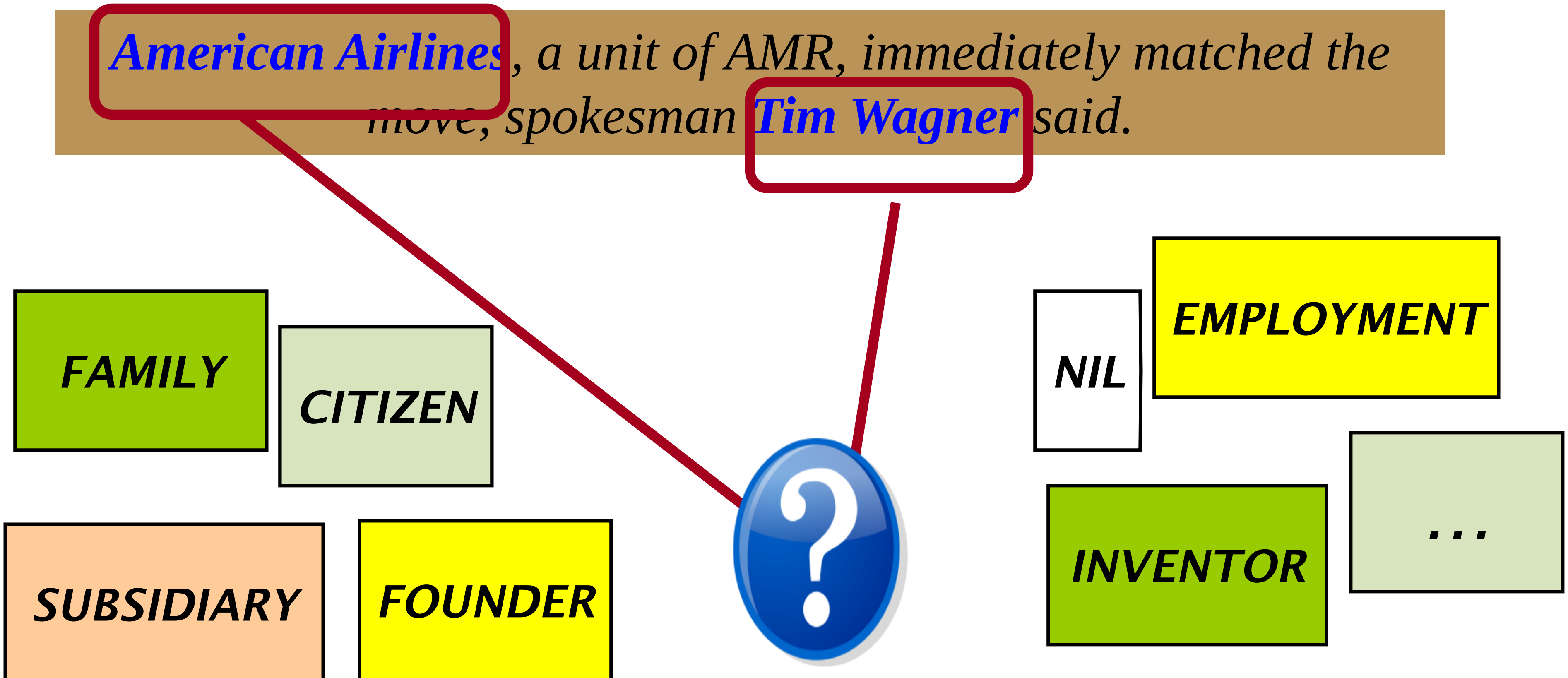
Automated Content Extraction (ACE)

17 sub-relations of 6 relations from 2008 "Relation Extraction Task"



Relation Extraction

Classify the relation between two entities in a sentence



Slide from D. Jurafsky

Word Features for Relation Extraction

Mention 1

American Airlines, *a unit of AMR, immediately matched the move, spokesman* ***Tim Wagner*** *said*

Mention 2

- Headwords of M1 and M2, and combination

  Airlines Wagner Airlines-Wagner

- Bag of words and bigrams in M1 and M2

  {American, Airlines, Tim, Wagner, American Airlines, Tim Wagner}

- Words or bigrams in particular positions left and right of M1/M2

  M2: -1 spokesman

  M2: +1 said

- Bag of words or bigrams between the two entities

  {a, AMR, of, immediately, matched, move, spokesman, the, unit}

Slide from Dan Jurafsky

Named Entity Type and Mention Level Features for Relation Extraction

Mention 1

American Airlines, *a unit of AMR, immediately matched the move, spokesman* ***Tim Wagner*** *said*

Mention 2

- Named-entity types
  - M1: ORG
  - M2: PERSON
- Concatenation of the two named-entity types
  - ORG-PERSON
- Entity Level of M1 and M2 (NAME, NOMINAL, PRONOUN)
  - M1: NAME [it or he would be PRONOUN]
  - M2: NAME [the company would be NOMINAL]

Slide from Dr. Jurafsky

Parse Features for Relation Extraction

Mention 1

American Airlines, *a unit of AMR, immediately matched the move, spokesman* ***Tim Wagner*** *said*

Mention 2

- Base syntactic chunk sequence from one to the other

  NP NP PP VP NP NP

- Constituent path through the tree from one to the other

  NP ↑ NP ↑ S ↑ S ↓ NP

- Dependency path

  Airlines matched Wagner said

Gazetteer and trigger word features for relation extraction

- Trigger list for family: kinship terms
  - parent, wife, husband, grandparent, etc. [from WordNet]
- Gazetteer:
  - Lists of useful geo or geopolitical words
    - Country name list
    - Other sub-entities

***American Airlines**, a unit of AMR, immediately matched the move, spokesman **Tim Wagner** said.*

Entity-based features	
Entity$_1$ type	ORG
Entity$_1$ head	*airlines*
Entity$_2$ type	PERS
Entity$_2$ head	*Wagner*
Concatenated types	ORGPERS
Word-based features	
Between-entity bag of words	{ *a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman* }
Word(s) before Entity$_1$	NONE
Word(s) after Entity$_2$	*said*
Syntactic features	
Constituent path	$NP \uparrow NP \uparrow S \uparrow S \downarrow NP$
Base syntactic chunk path	$NP \rightarrow NP \rightarrow PP \rightarrow NP \rightarrow VP \rightarrow NP \rightarrow NP$
Typed-dependency path	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

Slide from D. Jurafsky

Classifiers for supervised methods

- Now you can use any classifier you like
  - Decision Tree
  - MaxEnt
  - Naïve Bayes
  - SVM
  - ...
- Train it on the training set, tune on the dev set, test on the test set

Idea: label examples, train a classifier

You may also be thinking of Elon Musk (founder of SpaceX), who started PayPal.	✓
Elon Musk, co-founder of PayPal, went on to establish SpaceX, one of the most ...	✓
If Space Exploration (SpaceX), founded by Paypal pioneer Elon Musk succeeds, ...	✓
Entrepreneur Elon Musk announced the latest addition to the SpaceX arsenal ...	✗
Elon Musk tweeted Friday that SpaceX employees are “working on ventilators” ...	✗

founder	+1.56
founded	+1.41
establish	+1.23
co-founder	+1.01
PayPal	+0.35
announced	-0.23
addition	-0.32
tweeted	-0.44

Success! Better generalizability

Problem: labeling examples is expensive :-(

Summary: Supervised Relation Extraction

\+ Can get high accuracies with enough hand-labeled training data, if test similar enough to training

\- Labeling a large training set is expensive

\- Supervised models are brittle, don't generalize well to different genres

Slide from D. Jurafsky

Semi-Supervised Methods

- We'd like to minimize our reliance on having a large training set
- Instead, given a few examples or a few high-precision patterns, we'd like to generalize
  - This is sometimes referred to as "bootstrapping"

Relation Bootstrapping (Hearst 1992)

- Gather a set of seed pairs that have relation R
- Iterate:
  1. Find sentences with these pairs
  2. Look at the context between or around the pair and generalize the context to create patterns
  3. Use the patterns to grep for more pairs

Slide from D. Jurafsky

Bootstrapping

- <Mark Twain, Elmira> Seed tuple
  - Grep (google) for the environments of the seed tuple

  “Mark Twain is buried in Elmira, NY.”

  X is buried in Y

  “The grave of Mark Twain is in Elmira”

  The grave of X is in Y

  “Elmira is Mark Twain’s final resting place”

  Y is X’s final resting place.
- Use those patterns to grep for new tuples
- Iterate

Dipre: Extract <author, book> pairs

Brin, Sergei. 1998. Extracting Patterns and Relations from the World Wide Web.

- Start with 5 seeds:

Author	Book
Isaac Asimov	The Robots of Dawn
David Brin	Startide Rising
James Gleick	Chaos: Making a New Science
Charles Dickens	Great Expectations
William Shakespeare	The Comedy of Errors

- Find Instances:

  The Comedy of Errors, by William Shakespeare, was
  The Comedy of Errors, by William Shakespeare, is
  The Comedy of Errors, one of William Shakespeare's earliest attempts
  The Comedy of Errors, one of William Shakespeare's most

- Extract patterns (group by middle, take longest common prefix/suffix)

  `?x , by ?y ,`   `?x , one of ?y 's`

- Now iterate, finding new seeds that match the pattern

Slide from Dan Jurafsky

Snowball

- Similar iterative algorithm

Organization	Location of Headquarters
Microsoft	Redmond
Exxon	Irving
IBM	Armonk

- Group instances w/similar prefix, middle, suffix, extract patterns
  - But require that X and Y be named entities
  - And compute a confidence for each pattern

.69 *ORGANIZATION* `{'s, in, headquarters}` *LOCATION*

.75 *LOCATION* `{in, based}` *ORGANIZATION*

E. Agichtein and L. Gravano 2000. Snowball: Extracting Relations from Large Plain-Text Collections. ICDL

Distant Supervision

- Combine bootstrapping with supervised learning
  - Instead of 5 seeds,
    - Use a large database to get huge # of seed examples
  - Create lots of features from all these examples
  - Combine in a supervised classifier

Snow, Jurafsky, Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. NIPS 17
Fei Wu and Daniel S. Weld. 2007. Autonomously Semantifying Wikipeida. CIKM 2007
Mintz, Bills, Snow, Jurafsky. 2009. Distant supervision for relation extraction without labeled data. ACL09

Slide from D. Jurafsky

Idea: derive labels from an existing knowledge base (KB)

Assume sentences with related entities are positive examples

Assume sentences with unrelated entities are negative examples

relation	subject	object
founders	SpaceX	Elon_Musk
founders	Apple	Steve_Jobs
founders	Microsoft	Bill_Gates

Elon Musk, co-founder of PayPal, went on to establish SpaceX, one of the most ...	✓
Entrepreneur Elon Musk announced the latest addition to the SpaceX arsenal ...	✓
Elon Musk dismissed concerns that Apple was poaching the company's talent.	✗
Now we know what Apple would have done with Elon Musk if that deal had ...	✗

Hooray! Massive quantities of training data, practically free!

Qualm: are those assumptions reliable?

Distant supervision paradigm

- Like supervised classification:
  - Uses a classifier with lots of features
  - Supervised by detailed hand-created knowledge
  - Doesn't require iteratively expanding patterns
- Like unsupervised classification:
  - Uses very large amounts of unlabeled data
  - Not sensitive to genre issues in training corpus

Distantly supervised learning of relation extraction patterns

1. For each relation
2. For each tuple in big database
3. Find sentences in large corpus with both entities
4. Extract frequent features (parse, words, etc)
5. Train supervised classifier using thousands of patterns

Born-In

<Edwin Hubble, Marshfield>
<Albert Einstein, Ulm>

Hubble was born in Marshfield
Einstein, born (1879), Ulm
Hubble's birthplace in Marshfield

PER was born in LOC
PER, born (XXXX), LOC
PER's birthplace in LOC

$$P(\text{born-in} \mid f_1, f_2, f_3, \ldots, f_{70000})$$

Slide from D. Jurafsky

Distant supervision-Supervised Approach

Distant supervision is a powerful idea — but it has two limitations:

1. Not all sentences with related entities are truly positive examples

   Entrepreneur Elon Musk announced the latest addition to the SpaceX arsenal ... ✅ 🤥

   (but the benefit of *more* data outweighs the harm of noisier data)

2. Need an existing KB to start from — can't start from scratch

Unsupervised relation extraction

- Open Information Extraction:
  - extract relations from the web with no training data, no list of relations

1. Use parsed data to train a “trustworthy tuple” classifier
2. Single-pass extract all relations between NPs, keep if trustworthy
3. Assessor ranks relations based on text redundancy

(FCI, specializes in, software development)

(Tesla, invented, coil transformer)

M. Banko, M. Cararella, S. Soderland, M. Broadhead, and O. Etzioni. 2007. Open information extraction from the web. IJCAI

Slide from Dan Jurafsky

Evaluation of Supervised Relation Extraction

- Compute P/R/F_1 for each relation

$$P = \frac{\text{\# of correctly extracted relations}}{\text{Total \# of extracted relations}}$$

$$F_1 = \frac{2PR}{P+R}$$

$$R = \frac{\text{\# of correctly extracted relations}}{\text{Total \# of gold relations}}$$

Evaluation of Semi-supervised and Unsupervised Relation Extraction

- Since it extracts totally new relations from the web
  - There is no gold set of correct instances of relations!
    - Can't compute precision (don't know which ones are correct)
    - Can't compute recall (don't know which ones were missed)
- Instead, we can approximate precision (only)
  - Draw a random sample of relations from output, check precision manually

$$\hat{P} = \frac{\text{\# of correctly extracted relations in the sample}}{\text{Total \# of extracted relations in the sample}}$$

- Can also compute precision at different levels of recall.
  - Precision for top 1000 new relations, top 10,000 new relations, top 100,000
  - In each case taking a random sample of that set
- But no way to evaluate recall

Distant Supervision – Knowledge Bases

- One of the earliest attempts: entity and entity relations
- Hoffmann et al. (2010): Learn from Wikipedia infoboxes

Personal details	
Born	$\boxed{\text{November 19, 1949}}$ (age 72) New York City, New York, U.S.
Spouse(s)	$\boxed{\text{Michael W. Doyle}}$ (m. 1976)
Children	Abigail
Education	$\boxed{\text{Harvard University}}$ (BA, PhD) London School of Economics (MS)

Amy Gutmann was born on November 19, 1949,[2] in Brooklyn, New York,[2] the only child of Kurt and Beatrice Gutmann. ... She then entered Radcliffe College of Harvard University in 1967 on a scholarship as a math major with sophomore standing. ... She and her husband Michael Doyle have also funded an endowed undergraduate scholarship and an undergraduate research fund at Penn.

- Matching info box entities with context, to learn context-dependent relation extraction.
    - 5000+ relations
- Many follow-up work on de-noising, but with similar weak signals

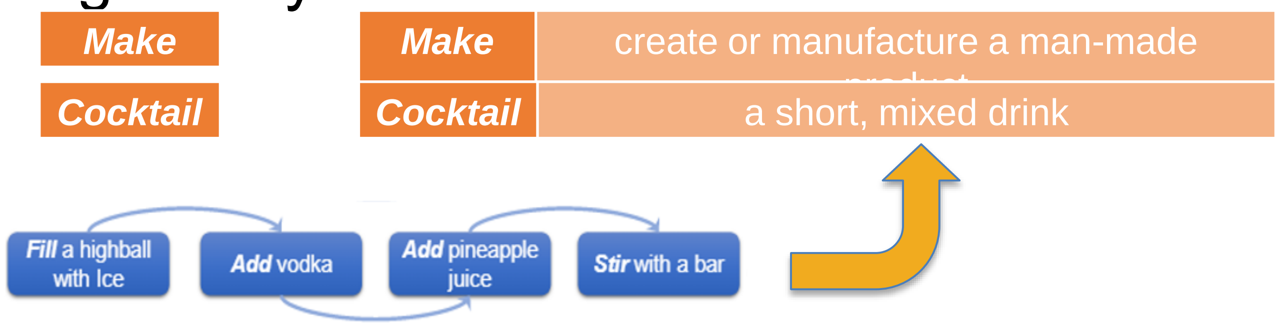
Hoffmann et al. Learning 5000 Relational Extractors. ACL 2010

Distant Supervision

- Mark joined **Amazon** a month ago.
  - What is the entity type?
- Weak Supervision:
  - From knowledge bases
    - Amazon.com, Inc is an American multinational technology company.
  - From weak but richer label representations
    - Word-embedding(company) is close to Word-embedding(Amazon)

Distant Supervision – Label Representations

- Chen et al. (2020): Event Process Typing
- Direct label understanding is difficult
  - Add glossary definition as a “weak” label defintion



Why using label glosses?

- *Semantically richer than labels themselves*
- *Capturing the association of a process-gloss pair (two sequences) is much easier*
- *Jump-starting few-shot label representations (and benefiting with fairer prediction)*

Chen et al. “What Are You Trying to Do?” Semantic Typing of Event Processes. CoNLL 2020

Distant Supervision Approaches

- Mark joined **Amazon** a month ago.
  - What is the entity type?
- Weak Supervision:
  - From knowledge bases
    - Amazon.com, Inc is an American multinational technology company.
  - From weak but richer label representations
    - Word-embedding(company) is close to Word-embedding(Amazon)
  - From pre-trained LMs
    - Amazon is a [MASK] <- [MASK] = company

Distant Supervision – Pretrained Language Models

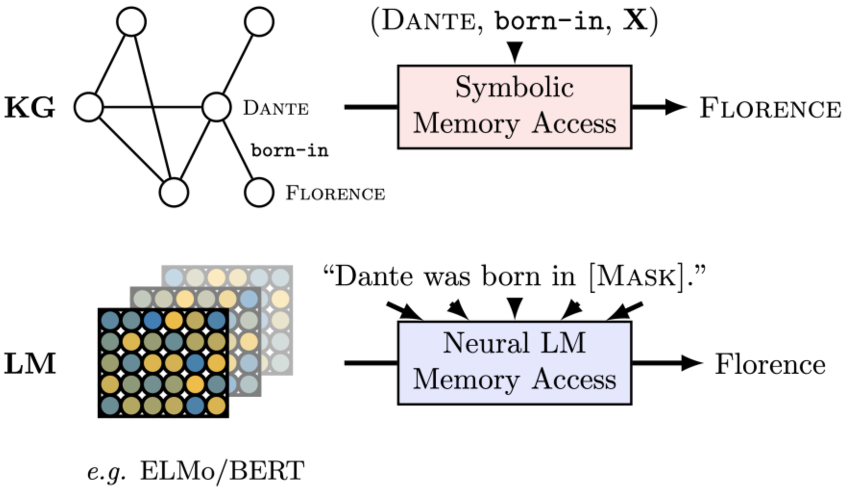
- Pre-trained language models can also be used as distant supervision
  - It did not use additional annotations
  - It is not task-specific
  - It contains inductive biases (weak signals)

Paris is a [MASK].

- PLMs are applied for IE in many creative ways
  - Contextual embeddings to replace word embeddings
  - Direct probing
  - Direct probing + task-specific finetuning

Distant Supervision – Pretrained Language Models

- Comparing to ELMo, BERT made direct probing easier
- Petroni et al. (2019): Language models as knowledge bases
  - Google-RE
  - 16.1% birth-place
  - 1.4% birth-date



Petroni et al. Language Models as Knowledge Bases?. EMNLP 2019

Distant Supervision – Pretrained Language Models

- Petroni et al. (2019): Language models as knowledge bases

	Relation	Query	Answer	Generation
T-Rex	P19	Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8], **Florence [-1.8]**, Naples [-1.9], Milan [-2.4], Bologna [-2.5]
	P20	Adolphe Adam died in ____.	Paris	**Paris [-0.5]**, London [-3.5], Vienna [-3.6], Berlin [-3.8], Brussels [-4.0]
	P279	English bulldog is a subclass of ____.	dog	dogs [-0.3], breeds [-2.2], **dog [-2.4]**, cattle [-4.3], sheep [-4.5]
	P37	The official language of Mauritius is ____.	English	**English [-0.6]**, French [-0.9], Arabic [-6.2], Tamil [-6.7], Malayalam [-7.0]
	P413	Patrick Oboya plays in ____ position.	midfielder	centre [-2.0], center [-2.2], **midfielder [-2.4]**, forward [-2.4], midfield [-2.7]
	P138	Hamburg Airport is named after ____.	Hamburg	Hess [-7.0], Hermann [-7.1], Schmidt [-7.1], **Hamburg [-7.5]**, Ludwig [-7.5]
	P364	The original language of Mon oncle Benjamin is ____.	French	**French [-0.2]**, Breton [-3.3], English [-3.8], Dutch [-4.2], German [-4.9]
	P54	Dani Alves plays with ____.	Barcelona	Santos [-2.4], Porto [-2.5], Sporting [-3.1], Brazil [-3.3], Portugal [-3.7]
	P106	Paul Toungui is a ____ by profession .	politician	lawyer [-1.1], journalist [-2.4], teacher [-2.7], doctor [-3.0], physician [-3.7]
	P527	Sodium sulfide consists of ____.	sodium	water [-1.2], sulfur [-1.7], **sodium [-2.5]**, zinc [-2.8], salt [-2.9]
	P102	Gordon Scholes is a member of the ____ political party.	Labor	Labour [-1.3], Conservative [-1.6], Green [-2.4], Liberal [-2.9], **Labor [-2.9]**
	P530	Kenya maintains diplomatic relations with ____.	Uganda	India [-3.0], **Uganda [-3.2]**, Tanzania [-3.5], China [-3.6], Pakistan [-3.6]
	P176	iPod Touch is produced by ____.	Apple	**Apple [-1.6]**, Nokia [-1.7], Sony [-2.0], Samsung [-2.6], Intel [-3.1]
	P30	Bailey Peninsula is located in ____.	Antarctica	**Antarctica [-1.4]**, Bermuda [-2.2], Newfoundland [-2.5], Alaska [-2.7], Canada [-3.1]
	P178	JDK is developed by ____.	Oracle	IBM [-2.0], Intel [-2.3], Microsoft [-2.5], HP [-3.4], Nokia [-3.5]
	P1412	Carl III used to communicate in ____.	Swedish	German [-1.6], Latin [-1.9], French [-2.4], English [-3.0], Spanish [-3.0]
	P17	Sunshine Coast, British Columbia is located in ____.	Canada	**Canada [-1.2]**, Alberta [-2.8], Yukon [-2.9], Labrador [-3.4], Victoria [-3.4]
	P39	Pope Clement VII has the position of ____ .	pope	cardinal [-2.4], Pope [-2.5], **pope [-2.6]**, President [-3.1], Chancellor [-3.2]
	P264	Joe Cocker is represented by music label ____.	Capitol	EMI [-2.6], BMG [-2.6], Universal [-2.8], **Capitol [-3.2]**, Columbia [-3.3]
	P276	London Jazz Festival is located in ____.	London	**London [-0.3]**, Greenwich [-3.2], Chelsea [-4.0], Camden [-4.6], Stratford [-4.8]
	P127	Border TV is owned by ____.	ITV	Sky [-3.1], **ITV [-3.3]**, Global [-3.4], Frontier [-4.1], Disney [-4.3]
	P103	The native language of Mammootty is ____.	Malayalam	**Malayalam [-0.2]**, Tamil [-2.1], Telugu [-4.8], English [-5.2], Hindi [-5.6]
	P495	The Sharon Cuneta Show was created in ____.	Philippines	Manila [-3.2], **Philippines [-3.6]**, February [-3.7], December [-3.8], Argentina [-4.0]
ConceptNet	AtLocation	You are likely to find a overflow in a ____.	drain	sewer [-3.1], canal [-3.2], toilet [-3.3], stream [-3.6], **drain [-3.6]**
	CapableOf	Ravens can ____.	fly	**fly [-1.5]**, fight [-1.8], kill [-2.2], die [-3.2], hunt [-3.4]
	CausesDesire	Joke would make you want to ____.	laugh	cry [-1.7], die [-1.7], **laugh [-2.0]**, vomit [-2.6], scream [-2.6]
	Causes	Sometimes virus causes ____.	infection	disease [-1.2], cancer [-2.0], **infection [-2.6]**, plague [-3.3], fever [-3.4]
	HasA	Birds have ____.	feathers	wings [-1.8], nests [-3.1], **feathers [-3.2]**, died [-3.7], eggs [-3.9]
	HasPrerequisite	Typing requires ____.	speed	patience [-3.5], precision [-3.6], registration [-3.8], accuracy [-4.0], **speed [-4.1]**
	HasProperty	Time is ____.	finite	short [-1.7], passing [-1.8], precious [-2.9], irrelevant [-3.2], gone [-4.0]
	MotivatedByGoal	You would celebrate because you are ____.	alive	happy [-2.4], human [-3.3], **alive [-3.3]**, young [-3.6], free [-3.9]
	ReceivesAction	Skills can be ____.	taught	acquired [-2.5], useful [-2.5], learned [-2.8], combined [-3.9], varied [-3.9]
	UsedFor	A pond is for ____.	fish	swimming [-1.3], fishing [-1.4], bathing [-2.0], **fish [-2.8]**, recreation [-3.1]

These predictions are highly relevant to typing and relation extraction

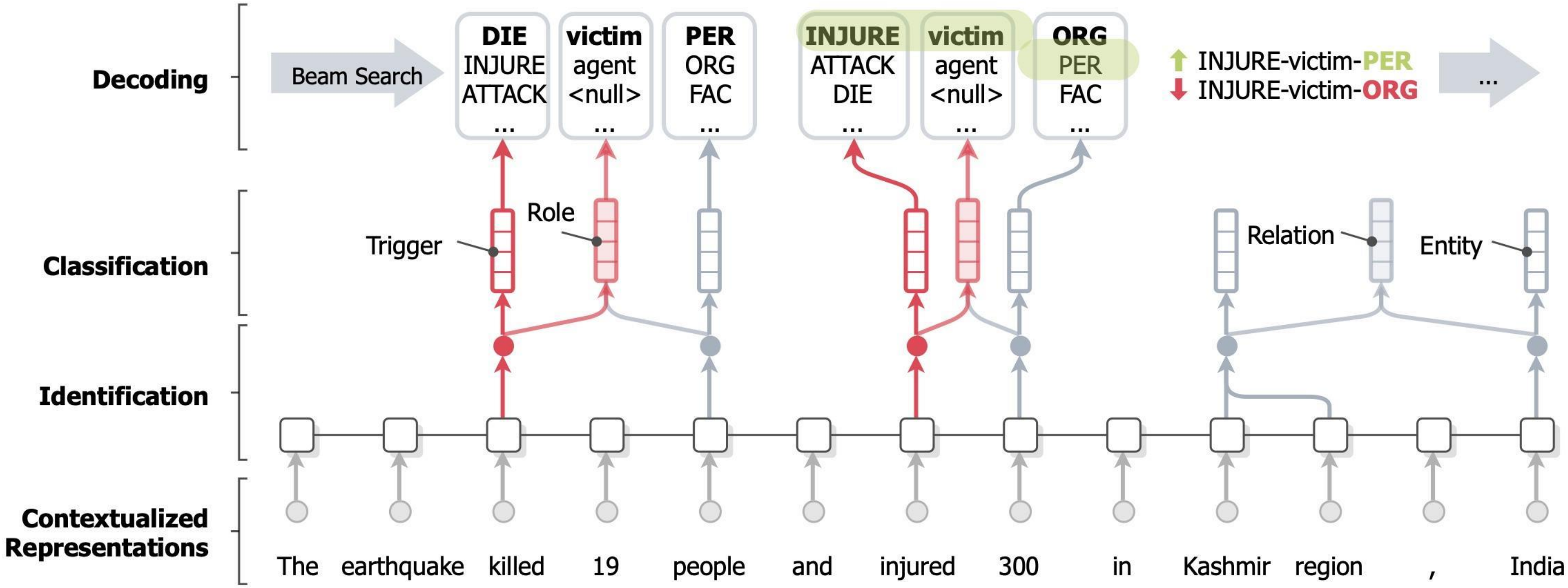
Petroni et al. Language Models as Knowledge Bases?. EMNLP 2019

- Mark joined **Amazon** a month ago.
  - What is the entity type?
- Weak Supervision:
  - From knowledge bases
    - Amazon.com, Inc is an American multinational technology company.
  - From weak but richer label representations
    - Word-embedding(company) is close to Word-embedding(Amazon)
  - From pre-trained LMs
    - Amazon is a [MASK] <- [MASK] = company.
  - From linguistic patterns
    - PER *join* company

- Zhou et al. (2020): Temporal Information Extraction from Patterns
  - Step 1: Extract distant signals of contextualized events and their duration, frequency etc. via linguistic patterns
  - Step 2: further pre-train a language model with extracted instances

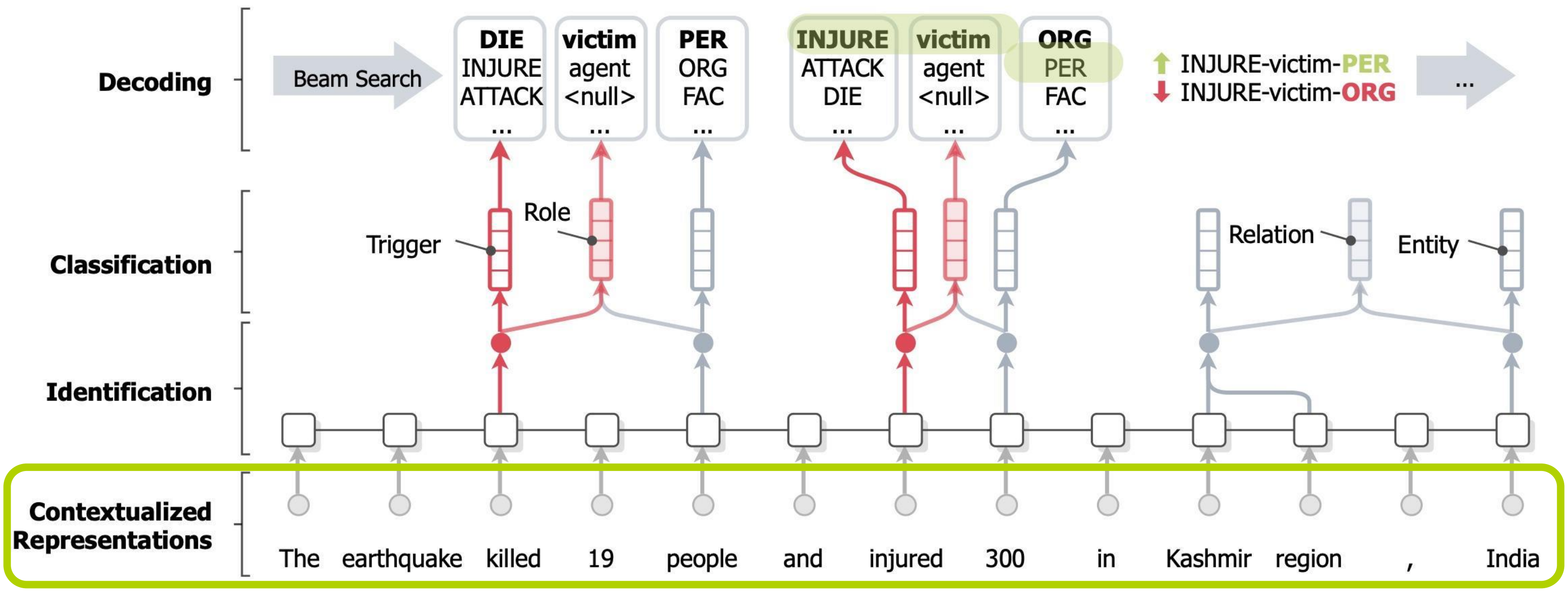
Zhou et al. Temporal Common Sense Acquisition with Minimal Supervision. ACL 2020

OneIE: An End-to-end Neural Model for IE



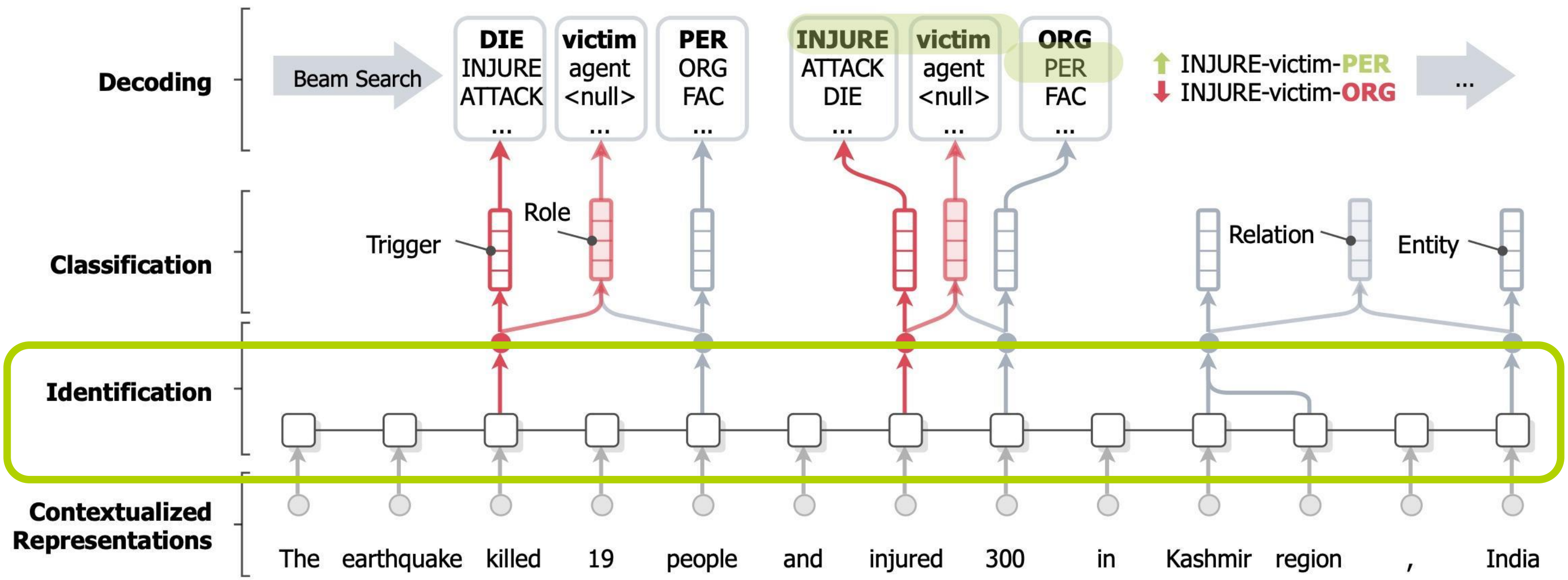
- Our OneIE framework extracts the information graph from a given sentence in four steps: encoding, identification, classification, and decoding

OneIE: An End-to-end Neural Model for IE



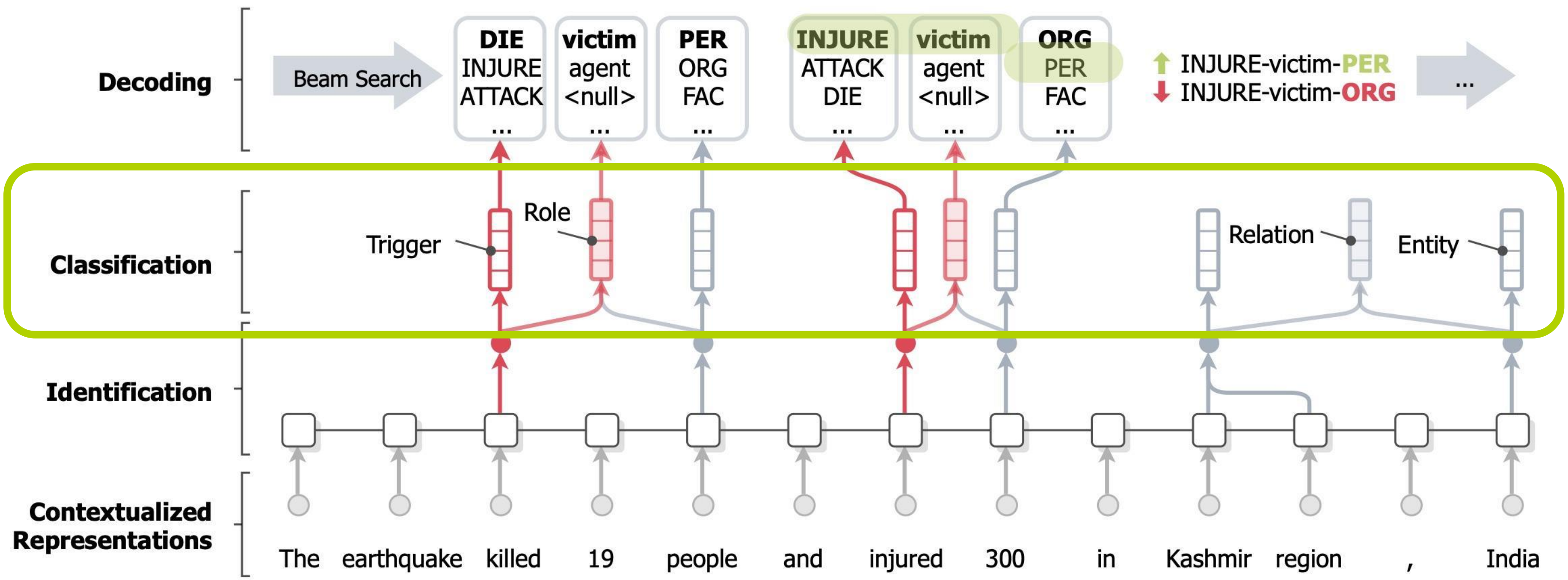
- **Encoding**: We use a BERT encoder to obtain the contextualized embedding of each token

OneIE: An End-to-end Neural Model for IE



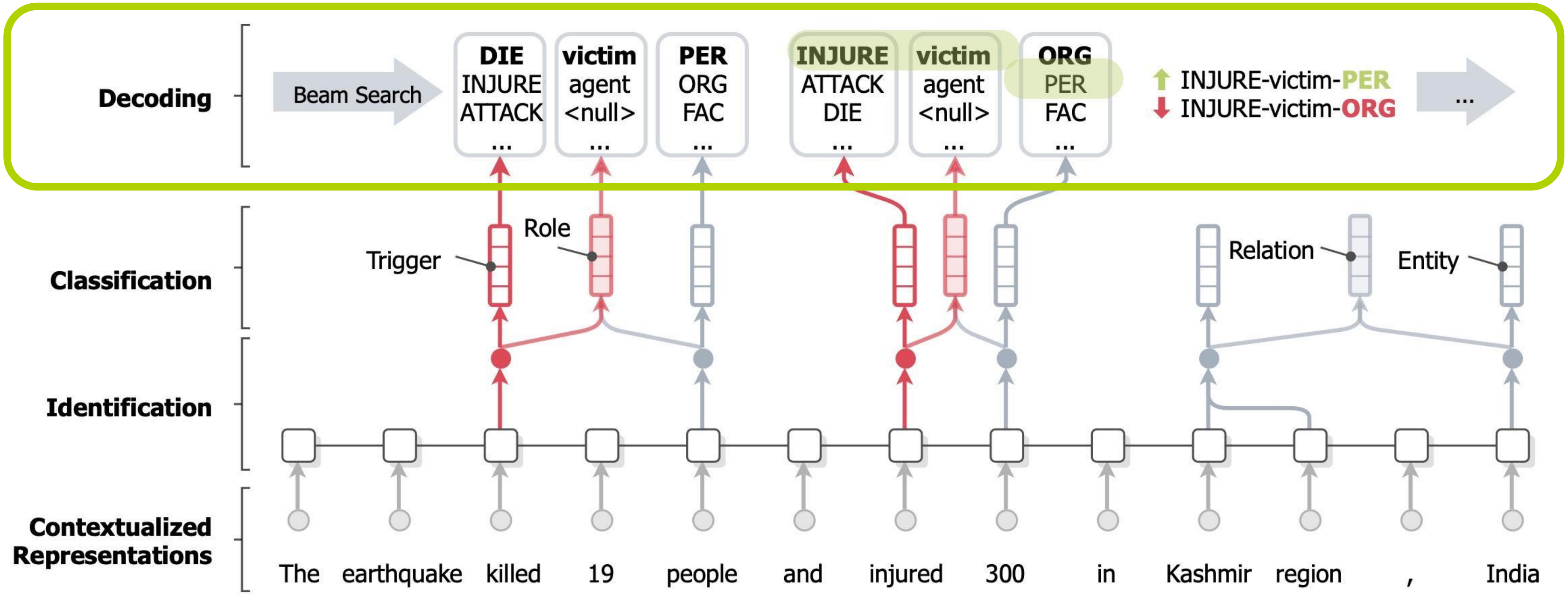
- **Identification**: We use CRF taggers to identify entity mentions and event triggers
- We define the identification loss as $\mathcal{L}^{\mathrm{I}} = -\log p(\boldsymbol{z}|\boldsymbol{X})$

OneIE: An End-to-end Neural Model for IE

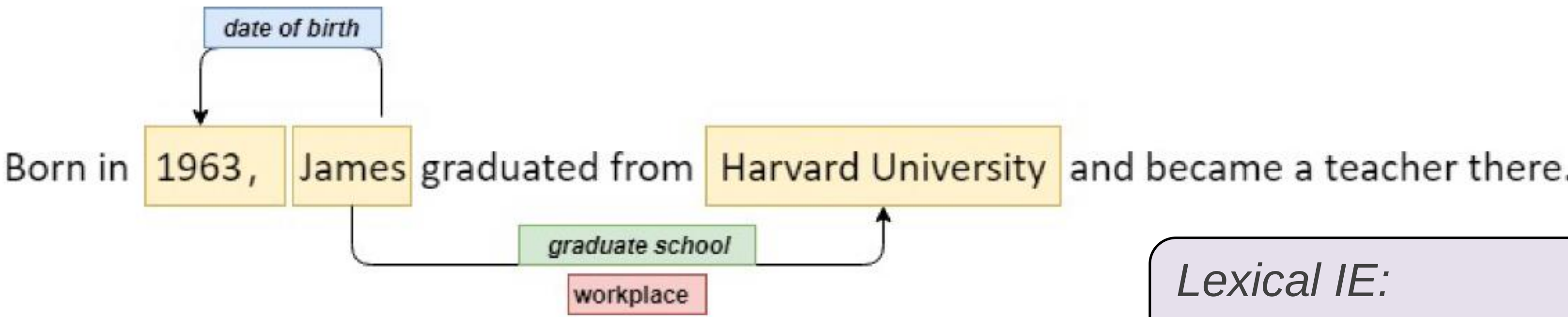


- **Classification**: We use task-specific feed-forward networks to calculate label scores for each node or edge
- We define the classification loss as $\mathcal{L}^t = -\frac{1}{N^t}\sum_{i=1}^{N^t} \boldsymbol{y}_i^t \log \hat{\boldsymbol{y}}_i^t$

OneIE: An End-to-end Neural Model for IE



- **Decoding**: In the test phase, we use a beam search decoder to find the information graph with the **highest global score**



- From the texts:
- 1 . Identify the concepts
  - Entities, events, terms, etc.
- 2. Identify the relations and other properties
  - Entity-entity / event-event
  - Temporal properties
  - etc.

Lexical IE:
- *Named entity recognition*
- *Entity/event typing*
- *Entity/event linking*

Relational IE:
- *Relation extraction*
  - *Entity / events*
  - *Sentence/Document*
  - *Temporal*
- *Coreference Resolution*

NLU Applications of Event Extraction

Narrative prediction

One day Wesley's auntie came over to visit. He was happy to see her, because he liked to play with her. When she started to give his little sister attention, he got **jealous**. He got **angry** at his auntie and **bit** his sister's hand when she wasn't looking.

Then what might happen?

O1: He was **scolded**.

O2: She **gave him a cookie** for being so nice.

jealous → angry → bit → scolded

jealous → angry → bit → get a cookie

Machine comprehension

Water is split, providing a source of electrons and protons (hydrogen ions, H^+) and giving off O_2 as a by-product. ***Light absorbed*** by chlorophyll drives a ***transfer of the electrons and hydrogen ions*** from water to an acceptor called $NADP^+$.

What can the splitting of water lead to?

A: Light absorption

B: Transfer of ions

Light absorbed → *Water is split* → *Transfer of the electrons and hydrogen ions*

Input Text

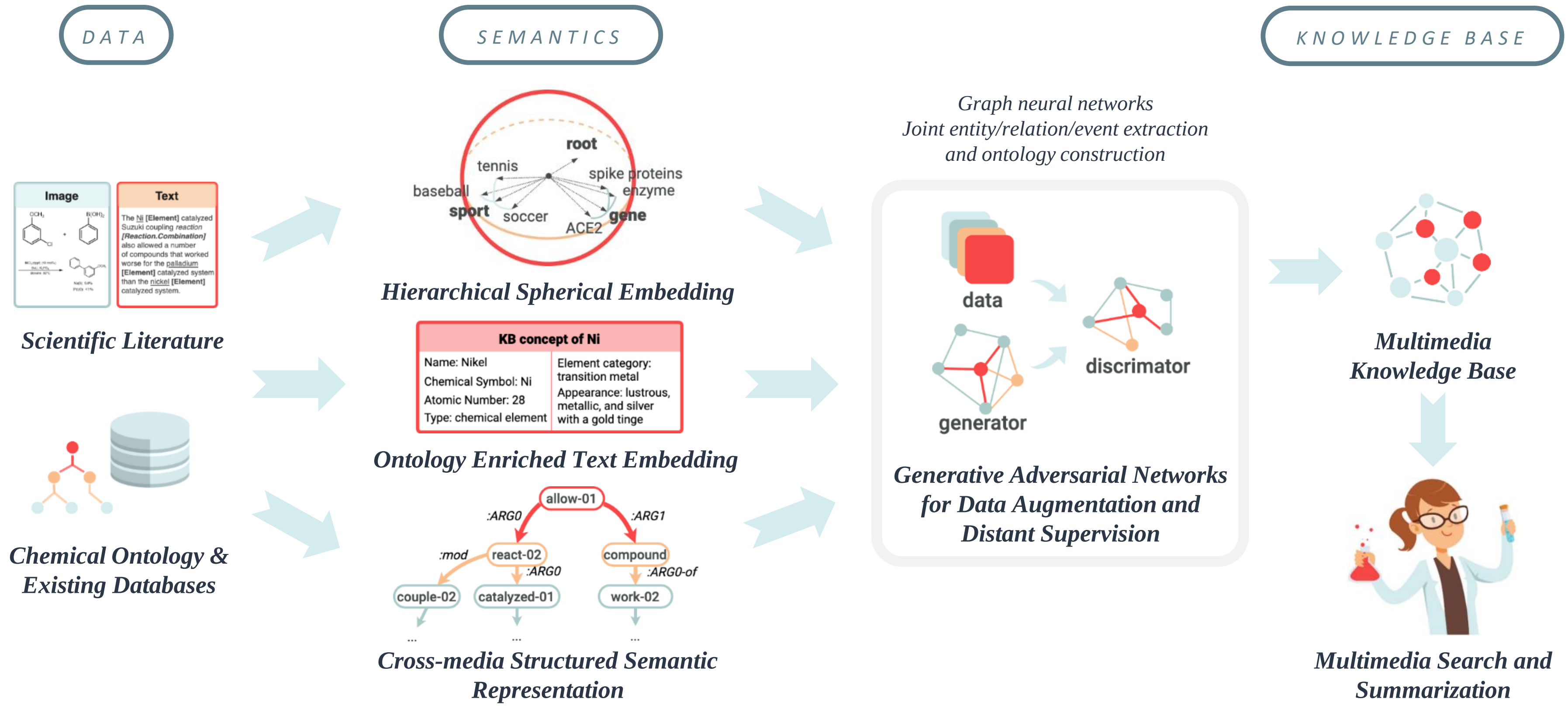
Unstructured text depends 100% on language understanding. *Semi-structured text* has some structure (layout) that can aid in understanding.

Unstructured Text	Semi-Structured Text
Professor John Skvoretz, U. of South Carolina, Columbia, will present a seminar entitled “Embedded Commitment,” on Thursday, May 4th from 4-5:30 in PH 223D.	Laura Petitte Department of Psychology McGill University Thursday, May 4, 1995 12:00 pm Baker Hall 355

innovate achieve lead

Name: Dr. Jeffrey D. Hermes
Affiliation: Department of AutoImmune Diseases
Research & Biophysical Chemistry Merch Research Laboratories
Title: "MHC Class II: A Target for Specific
Immunomodulation of the Immune Response"
Host/e-mail: Robert Murphy
Date: Wednesday, May 3, 1995
Time: 3:30 p.m.
Place: Mellon Institute Conference Room
Sponsor: MERCK RESEARCH LABORATORIES

GOAL: converting unstructured DATA to structured KNOWLEDGE



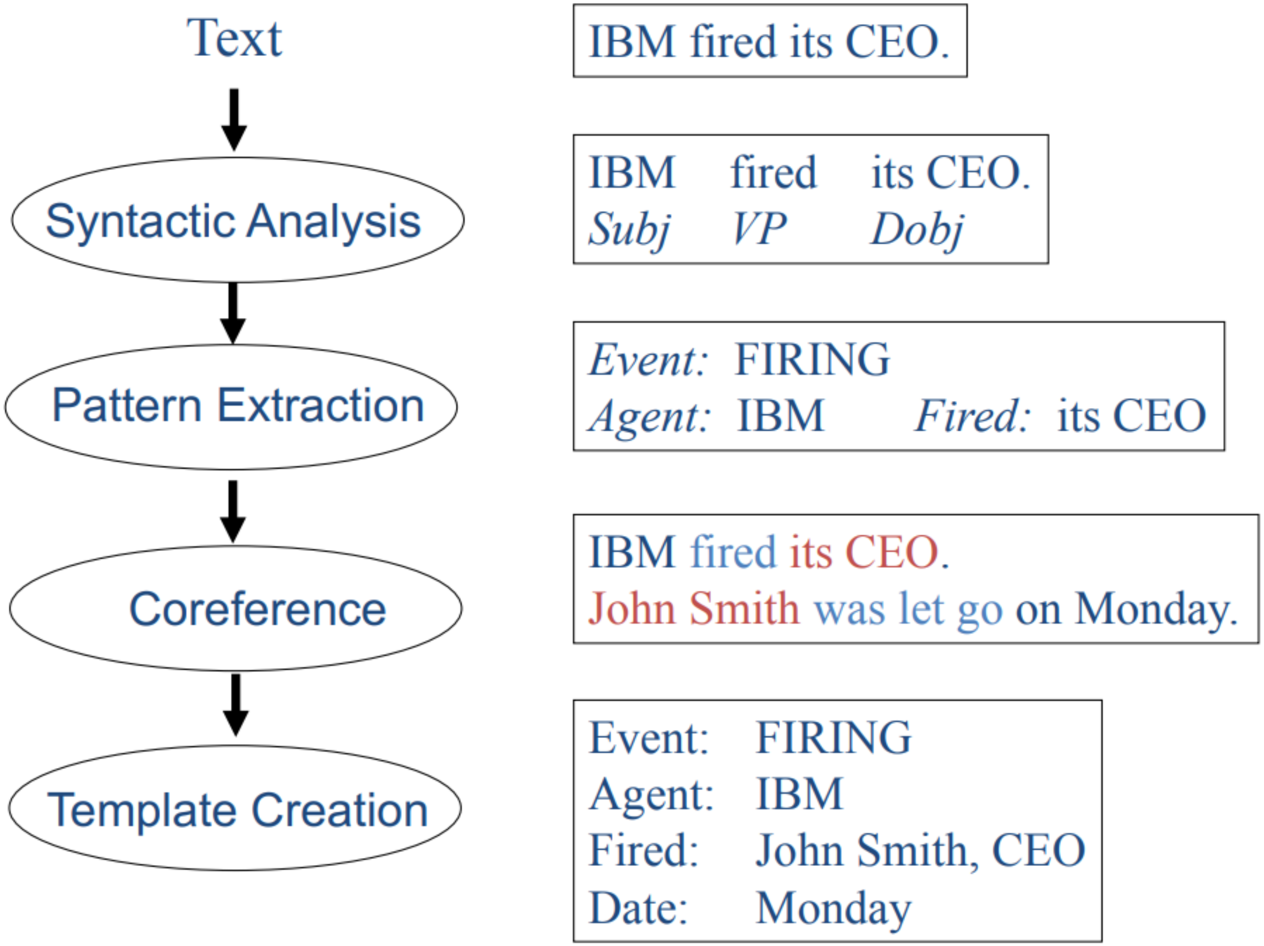
Event mention Extraction

- An event is specific occurrence that implies a change of states
- **event trigger**: the main word which most clearly expresses an event occurrence
- **event arguments**: the mentions that are involved in an event (participants)
- **event mention**: a phrase or sentence within which an event is described, including trigger and arguments
- Automatic Content Extraction defined 8 types of events, with 33 subtypes

Argument, role=victim   trigger

ACE event type/subtype	*Event Mention Example*
Life/Die	**Kurt Schork** died in **Sierra Leone yesterday**
Transaction/Transfer	**GM** sold **the company** in **Nov 1998** to **LLC**
Movement/Transport	**Homeless people** have been moved to **schools**
Business/Start-Org	**Schweitzer** founded **a hospital** in **1913**
Conflict/Attack	the attack on **Gaza** killed **13**
Contact/Meet	**Arafat's cabinet** met for **4 hours**
Personnel/Start-Position	**She** later recruited **the nursing student**
Justice/Arrest	**Faison** was wrongly arrested on suspicion of murder

Pattern-based Template-Filling Pipeline



Supervised Learning Approach

- Build a classifier as a sequence tagging model.
- Each document is processed sequentially and each token is labeled as Extraction or Non-Extraction.
- Ex: B (beginning), I (inside), or O (outside) tags.
- Features are usually simple: e.g., words, POS tags, orthography, and a small context window of preceding/following words

Supervised Event Mention Extraction: Methods

- Staged classifiers
  - Trigger Classifier
    - to distinguish event instances from non-events, to classify event instances by type
  - Argument Classifier
    - to distinguish arguments from non-arguments
  - Role Classifier
    - to classify arguments by argument role
  - Reportable-Event Classifier
    - to determine whether there is a reportable event instance
- Can choose any supervised learning methods such as MaxEnt and SVMs

(Ji and Grishman, 2008)

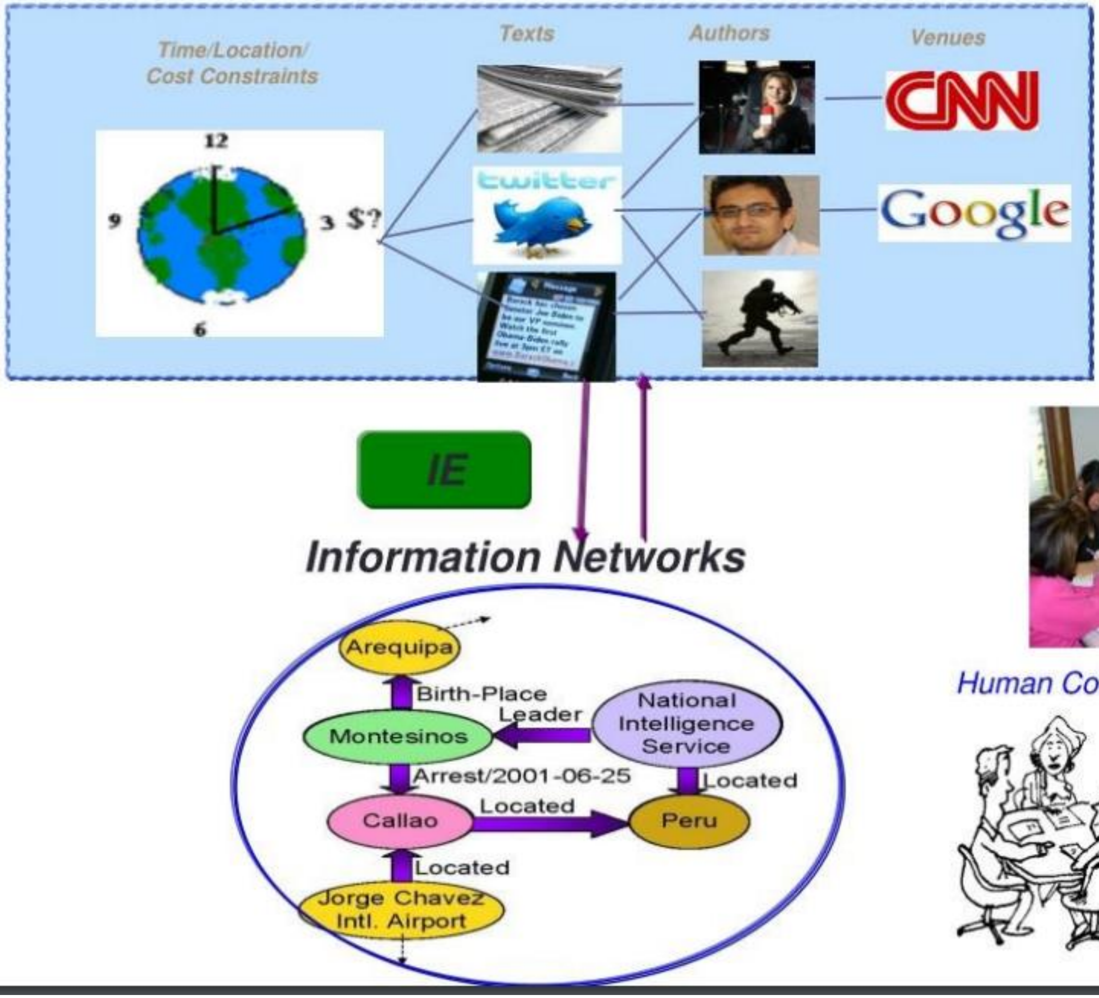
Typical Event Mention Extraction Features

- Trigger Labeling
  - Lexical
    - Tokens and POS tags of candidate trigger and context words
  - Dictionaries
    - Trigger list, synonym gazetteers
  - Syntactic
    - the depth of the trigger in the parse tree
    - the path from the node of the trigger to the root in the parse tree
    - the phrase structure expanded by the parent node of the trigger
    - the phrase type of the trigger
  - Entity
    - the entity type of the syntactically nearest entity to the trigger in the parse tree
    - the entity type of the physically nearest entity to the trigger in the sentence

- Argument Labeling
  - Event type and trigger
    - Trigger tokens
    - Event type and subtype
  - Entity
    - Entity type and subtype
    - Head word of the entity mention
  - Context
    - Context words of the argument candidate
  - Syntactic
    - the phrase structure expanding the parent of the trigger
    - the relative position of the entity regarding to the trigger (before or after)
    - the minimal path from the entity to the trigger
    - the shortest length from the entity to the trigger in the parse tree

(Chen and Ji, 2009)

IE in Rich Contexts

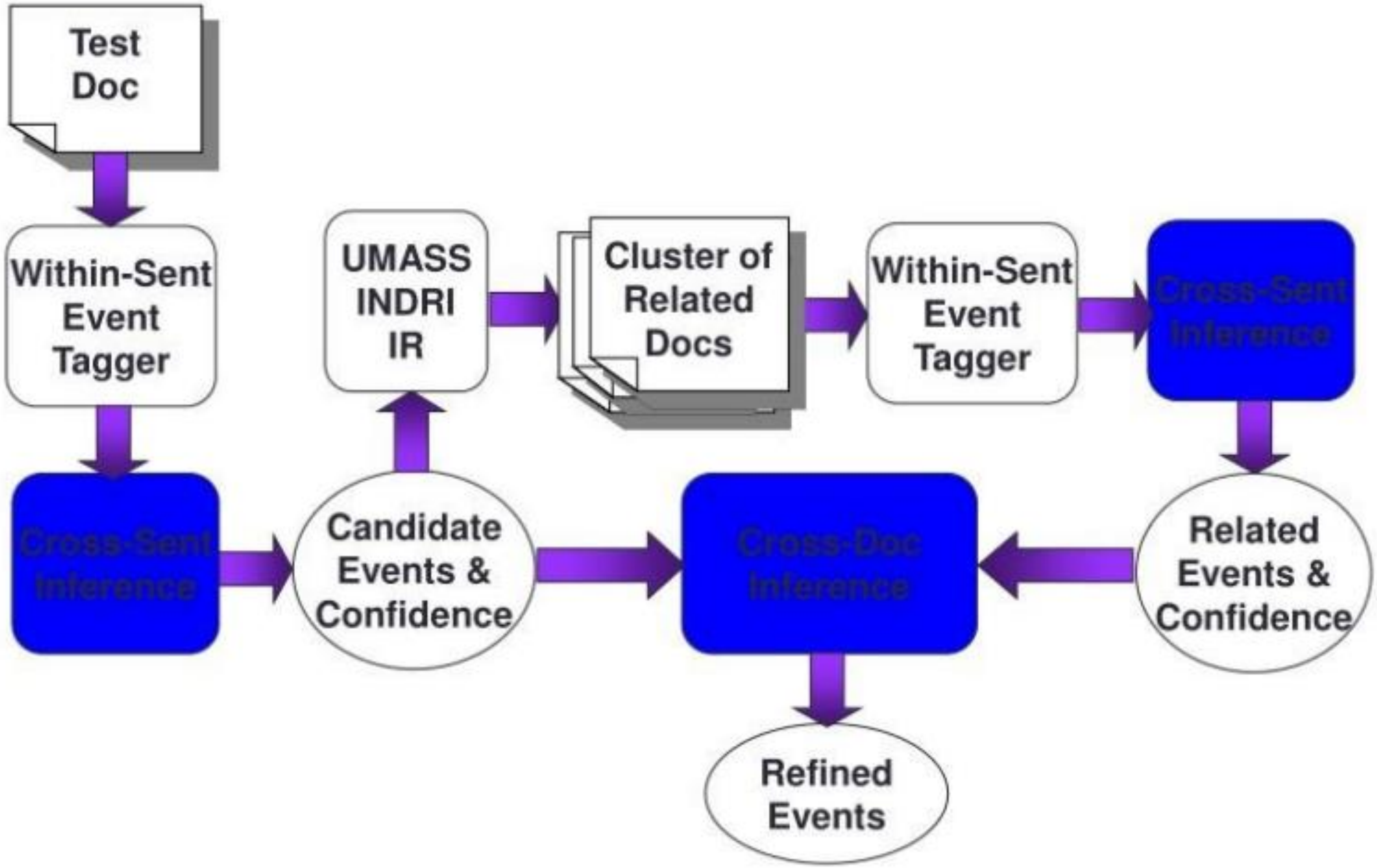


Human Collaborative Learning

Capture Information Redundancy

- When the data grows beyond some certain size, IE task is naturally embedded in rich contexts; the extracted facts become inter-dependent
- Leverage Information Redundancy from:
  - Large Scale Data (Chen and Ji, 2011)
  - Background Knowledge (Chan and Roth, 2010; Rahman and Ng, 2011)
  - Inter-connected facts (Li and Ji, 2011; Li et al., 2011; e.g. Roth and Yih, 2004; Gupta and Ji, 2009; Liao and Grishman, 2010; Hong et al., 2011)
  - Diverse Documents (Downey et al., 2005; Yangarber, 2006; Patwardhan and Riloff, 2009; Mann, 2007; Ji and Grishman, 2008)
  - Diverse Systems (Tamang and Ji, 2011)
  - Diverse Languages (Snover et al., 2011)
  - Diverse Data Modalities (text, image, speech, video…)
- But how? Such knowledge might be overwhelming…

Cross Sentences/Doc Event Inferencing



Within Sentence Extraction

1. Pattern matching
   - Build a pattern from each ACE training example of an event
     - British and US forces reported gains in the advance on Baghdad

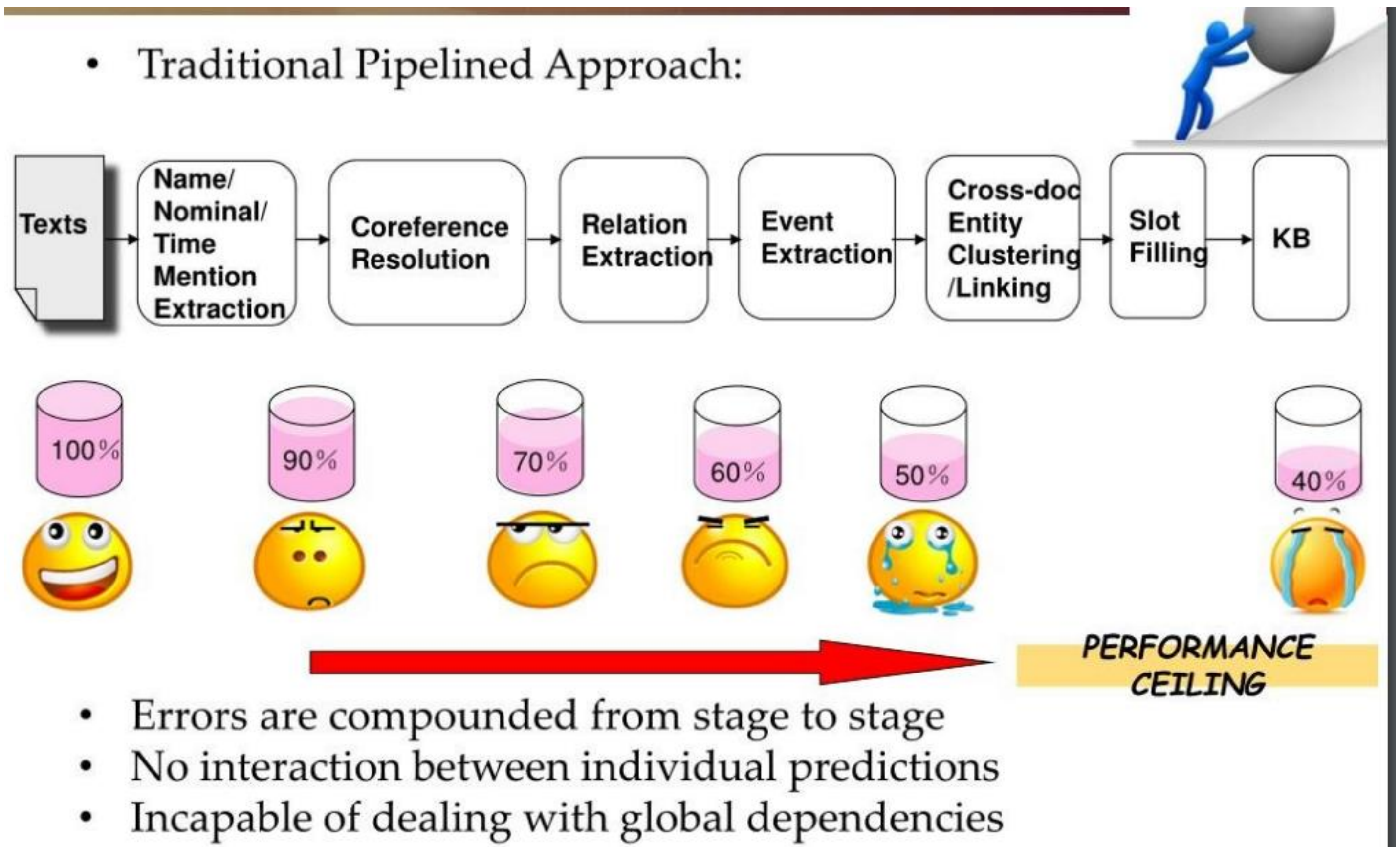
       → PER report gain in advance on LOC
2. MaxEnt models
   ① Trigger Classifier
     - to distinguish event instances from non-events, to classify event instances by type

   ② Argument Classifier
     - to distinguish arguments from non-arguments

   ③ Role Classifier
     - to classify arguments by argument role

   ④ Reportable-Event Classifier
     - to determine whether there is a reportable event instance

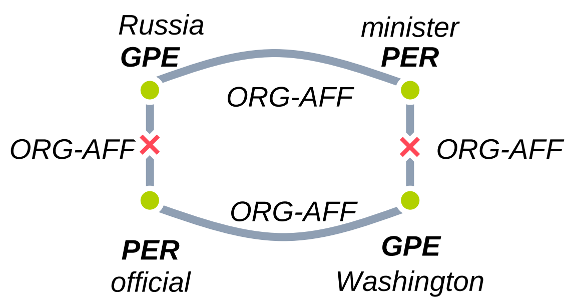
Event Mention Extraction



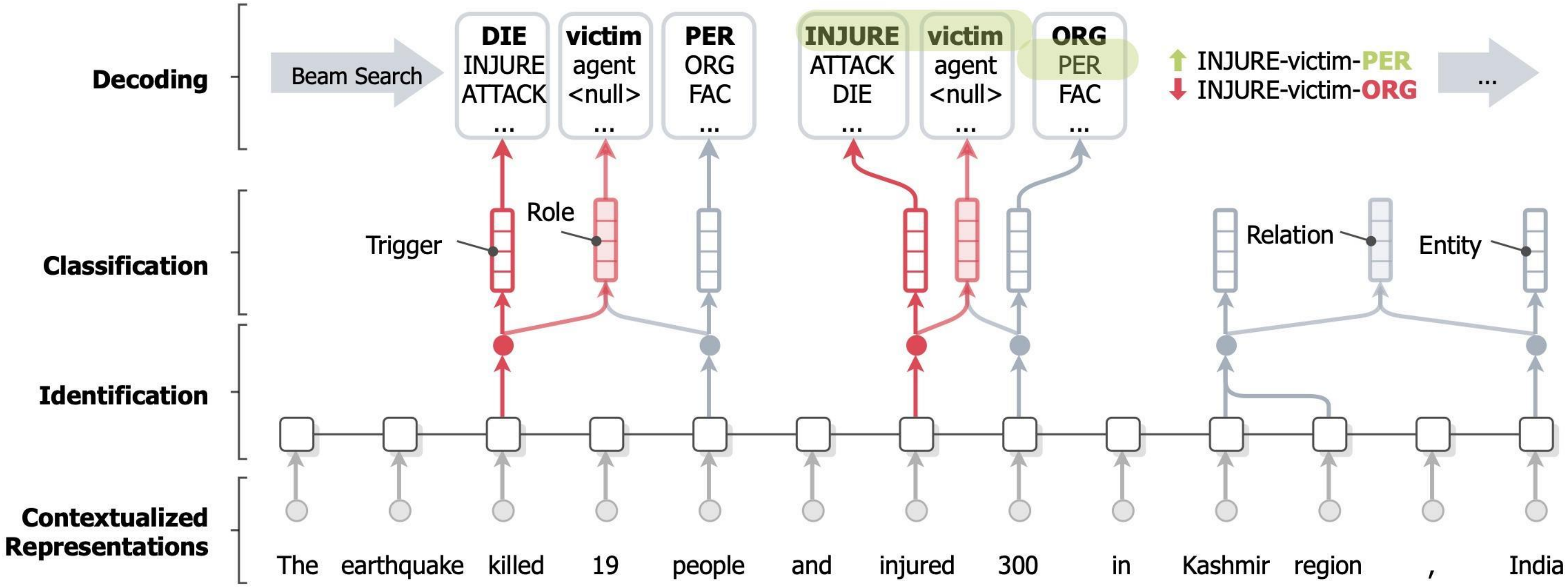
Motivation

- Pipeline models suffer from the **error propagation problem** and disallow interactions among components.
- Existing neural models do not explicitly model **cross-subtask and cross-instance interactions** among knowledge elements.

Russia's *foreign* ***minister*** *expressed outrage at suggestions from a top* ***Washington official*** *last week...*

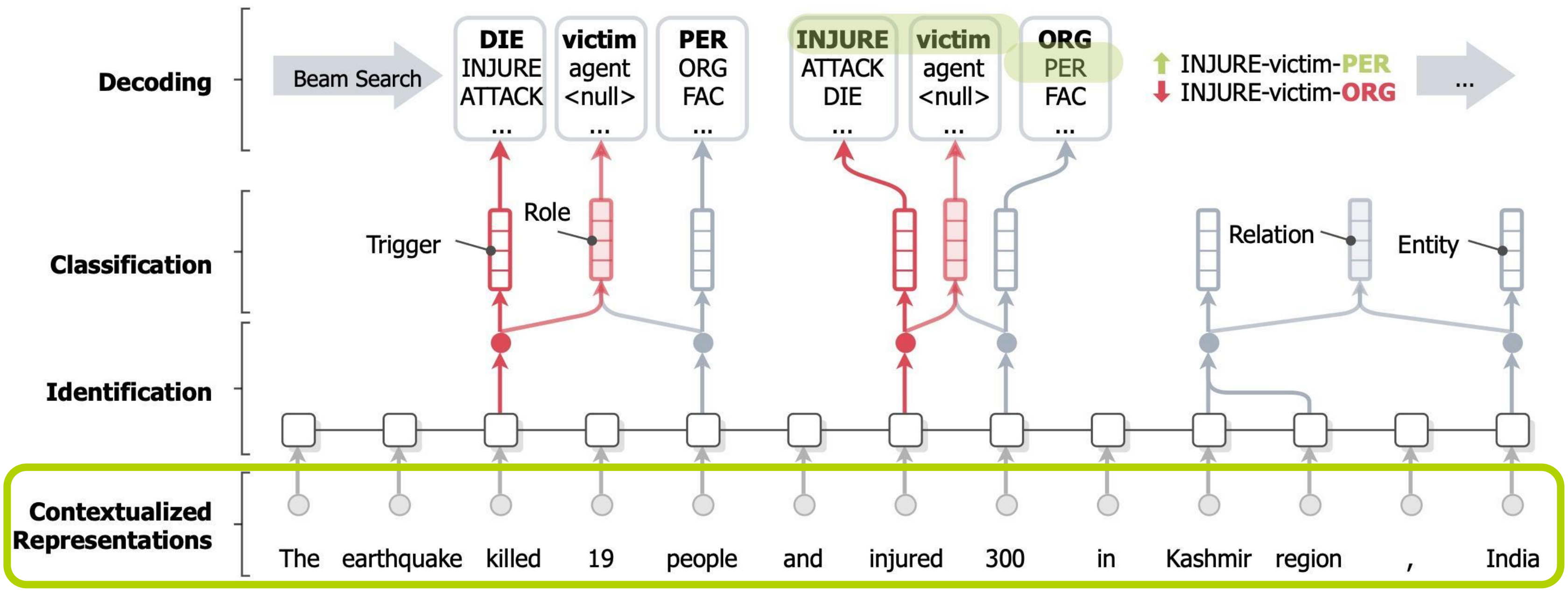


OneIE: An End-to-end Neural Model for IE



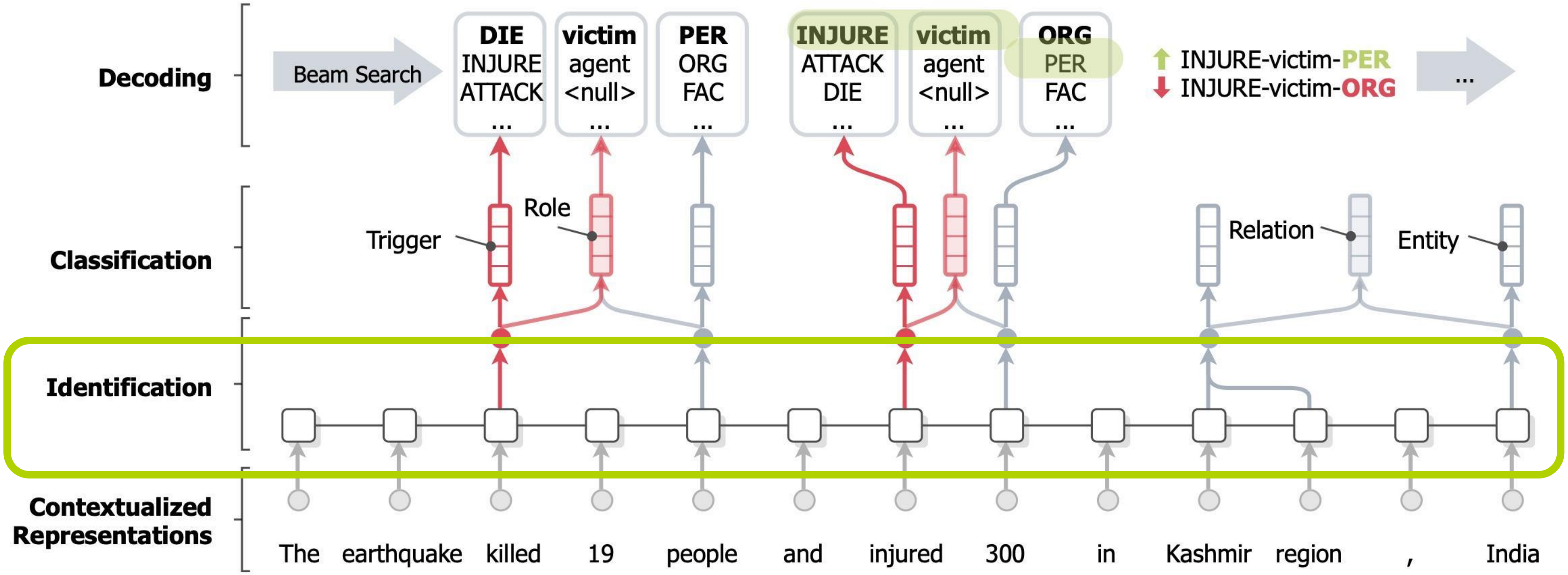
- Our OneIE framework extracts the information graph from a given sentence in four steps: encoding, identification, classification, and decoding

OneIE: An End-to-end Neural Model for IE



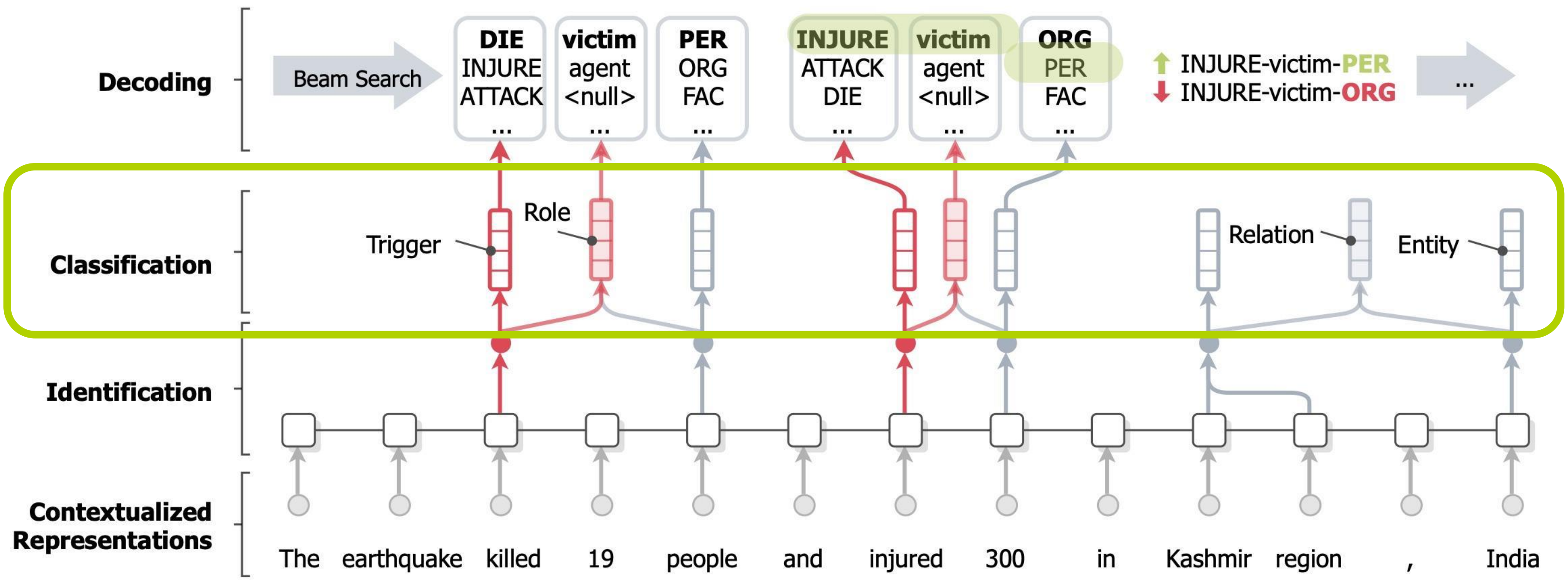
- **Encoding**: We use a BERT encoder to obtain the contextualized embedding of each token

OneIE: An End-to-end Neural Model for IE



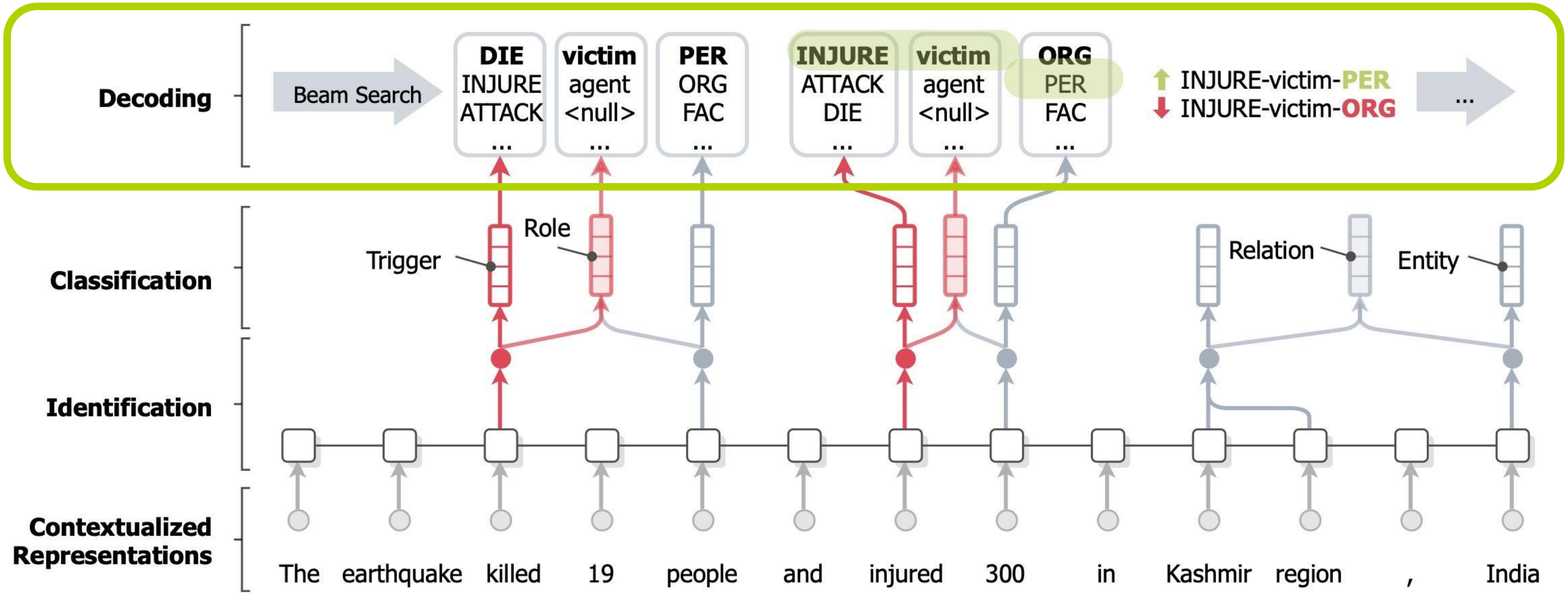
- **Identification**: We use CRF taggers to identify entity mentions and event triggers
- We define the identification loss as $\mathcal{L}^{\mathrm{I}} = -\log p(\boldsymbol{z}|\boldsymbol{X})$

OneIE: An End-to-end Neural Model for IE



- **Classification**: We use task-specific feed-forward networks to calculate label scores for each node or edge
- We define the classification loss as $\mathcal{L}^{t} = -\frac{1}{N^t}\sum_{i=1}^{N^t} \boldsymbol{y}_i^t \log \hat{\boldsymbol{y}}_i^t$

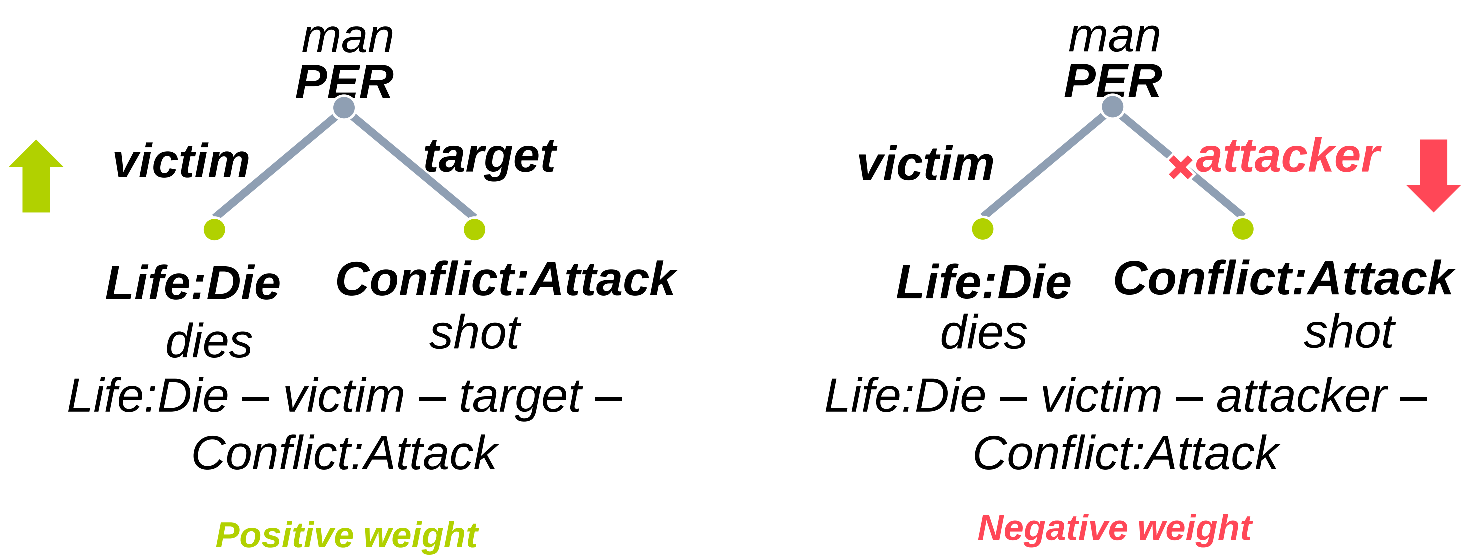
OneIE: An End-to-end Neural Model for IE



- **Decoding**: In the test phase, we use a beam search decoder to find the information graph with the **highest global score**

Incorporating Global Features

- We design a set of *global feature templates* (e.g., event_type_1 – role_1 – role_2 – event_type_2 : an entity acts a role_1 argument for an event_type_1 event and a role_2 argument for an event_type_2 event in the same sentence)
- The model learns the *weight* of each feature during training



Salient Global Features

- Salient positive and negative global features learned by the model
- Global features are explainable

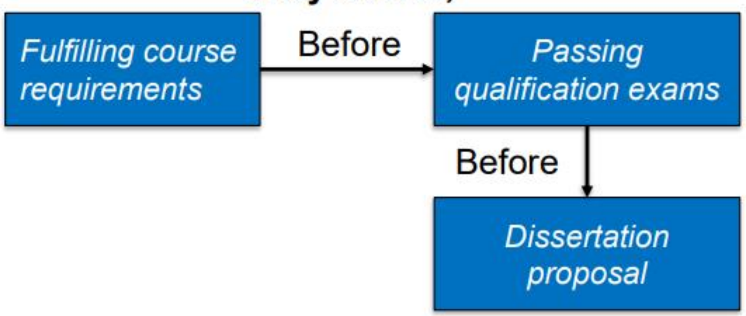
	Features	Weight
1	A **Transport** event has only one **Destination** argument	2.61
2	An **Attack** event has only one **Place** argument	2.31
3	A **PER-SOC** relation exists between two **PER** entities	1.51
4	A **Beneficiary** argument is a **PER** entity	0.93
5	An entity has an **ORG-AFF** relation with multiple entities	-3.21
6	An event has two **Place** arguments	-2.47
7	A **Transport** event has multiple **Destination** argument	-2.25
8	An entity has a **GEN-AFF** relation with multiple entities	-2.02

Events Process

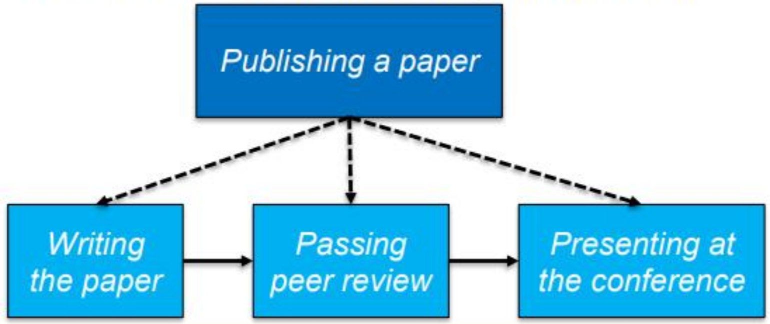
Extraction only is not enough.

Events are **NOT simple, static** predicates.

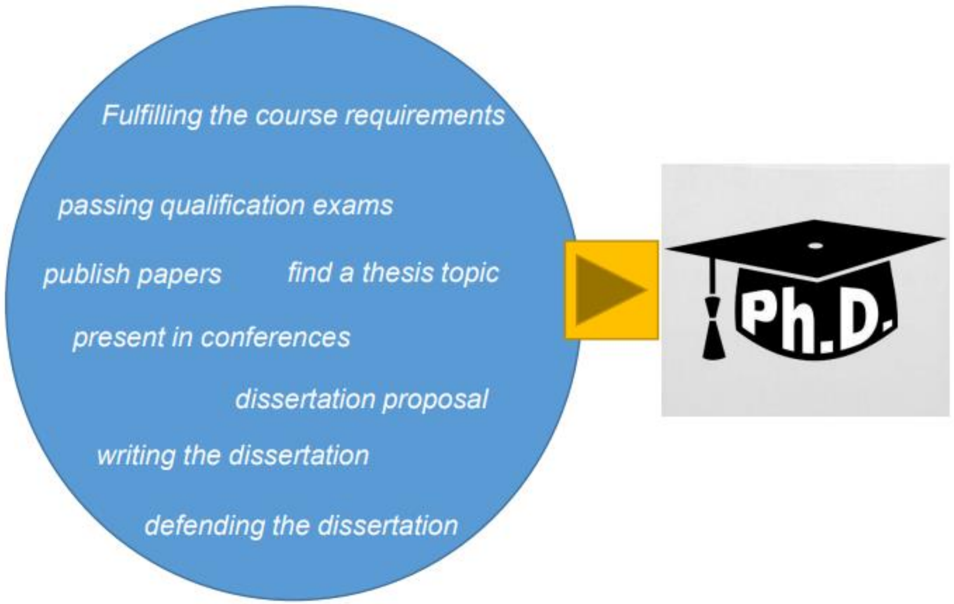
They evolve,



are described in different granularities,



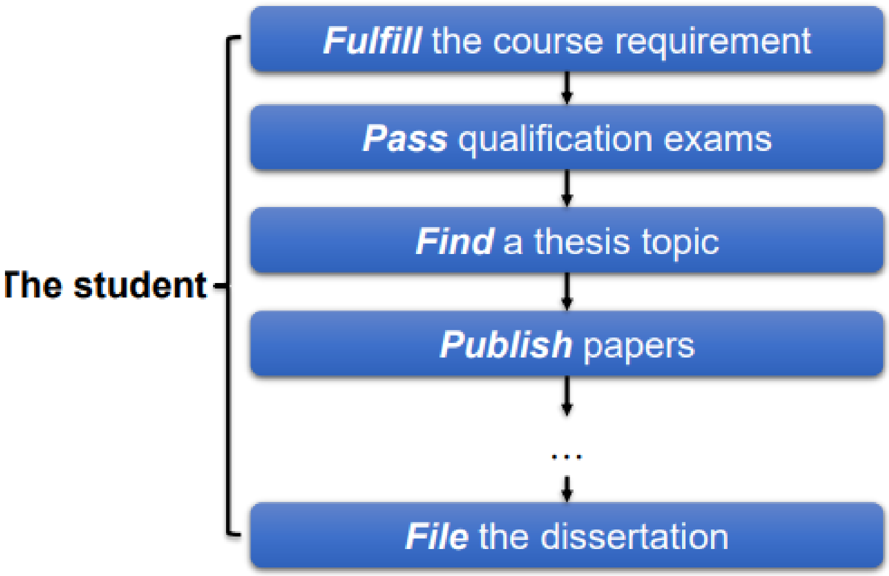
and are always directed by specific intents or central goals [Zacks et al. *Nature Neuroscience*, 2001]



Event Process

An event process (or event chain)

- Partially ordered events that are centered around common protagonists [Chambers et al., ACL-08]



Prediction problems on event processes

Event process completion

- What happens next?

Intention prediction

- What is the goal of "*digging a hole, putting some seeds in the hole* and *filling it with soil*"?

Membership prediction

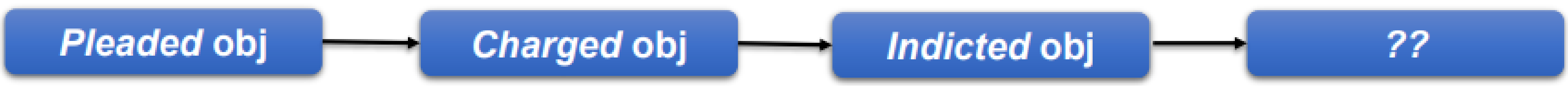
- What are the steps of "buying a car"?

Salience prediction

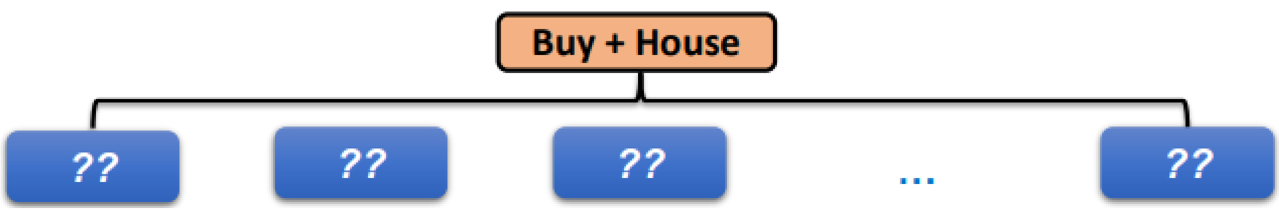
- Is *defending the dissertation* more important than *doing an internship*?

Predicting Events

1. Predicting steps of the process



2. Inducing the entire process from scratch.



Predicting Events

Chambers and Jurafsky. Unsupervised Learning of Narrative Event Chains. ACL-08

Unsupervised event process completion can be done using corpus statistics (Gigaword in this work)

- Capturing the co-occurrence of events using pointwise mutual information

$$pmi(e(w,d), e(v,g))$$

- The next most likely forthcoming event can be found by maximizing the accumulated PMI

$$\max_{j:0<j<m} \sum_{i=0}^{n} pmi(e_i, f_j)$$

(*n*: #events in the process; *m*: #events in the vocabulary.

Known events:
(pleaded subj), (admits subj), (convicted obj)

Likely Events:

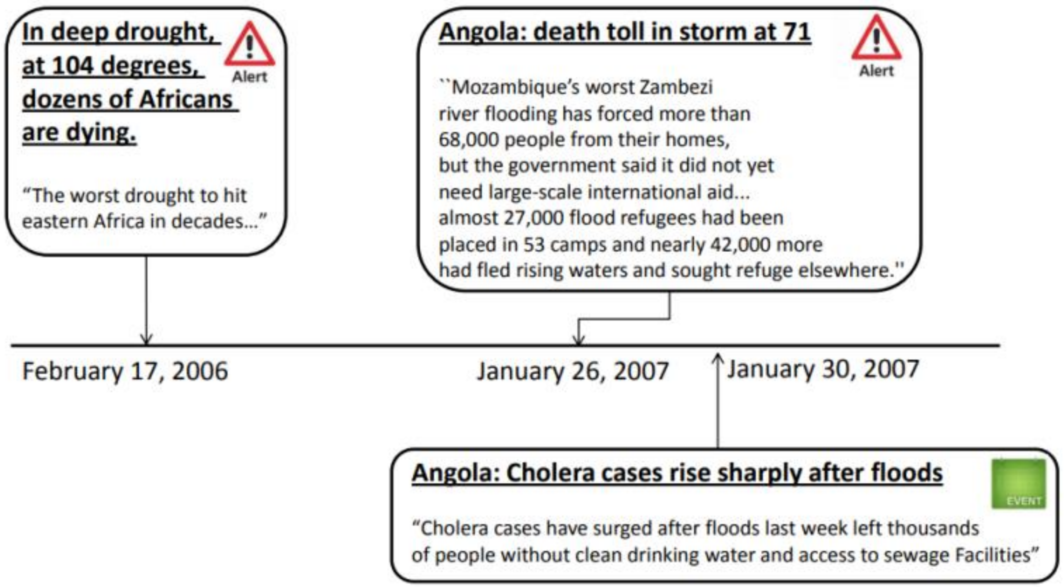
sentenced obj	0.89	indicted obj	0.74
paroled obj	0.76	fined obj	0.73
fired obj	0.75	denied subj	0.73

Improves narrative cloze tests (36% improvement on NYT Narrative Cloze).

Predicting Events

Radinsky and Horvitz. Mining the Web to Predict Future Events. WSDM, 2013

Extension of the event chain model on multiple **dated** and **topically cohesive** documents.

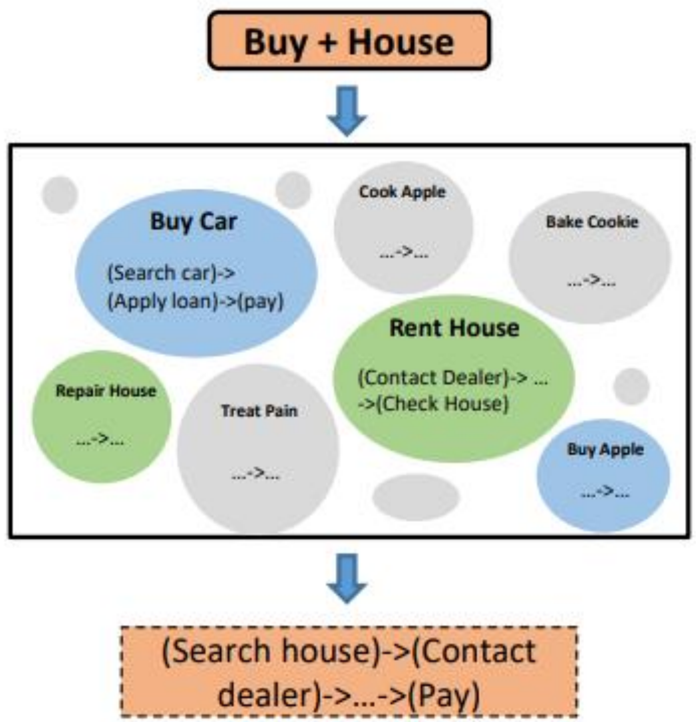


Maximum-entropy-based event co-occurence model.

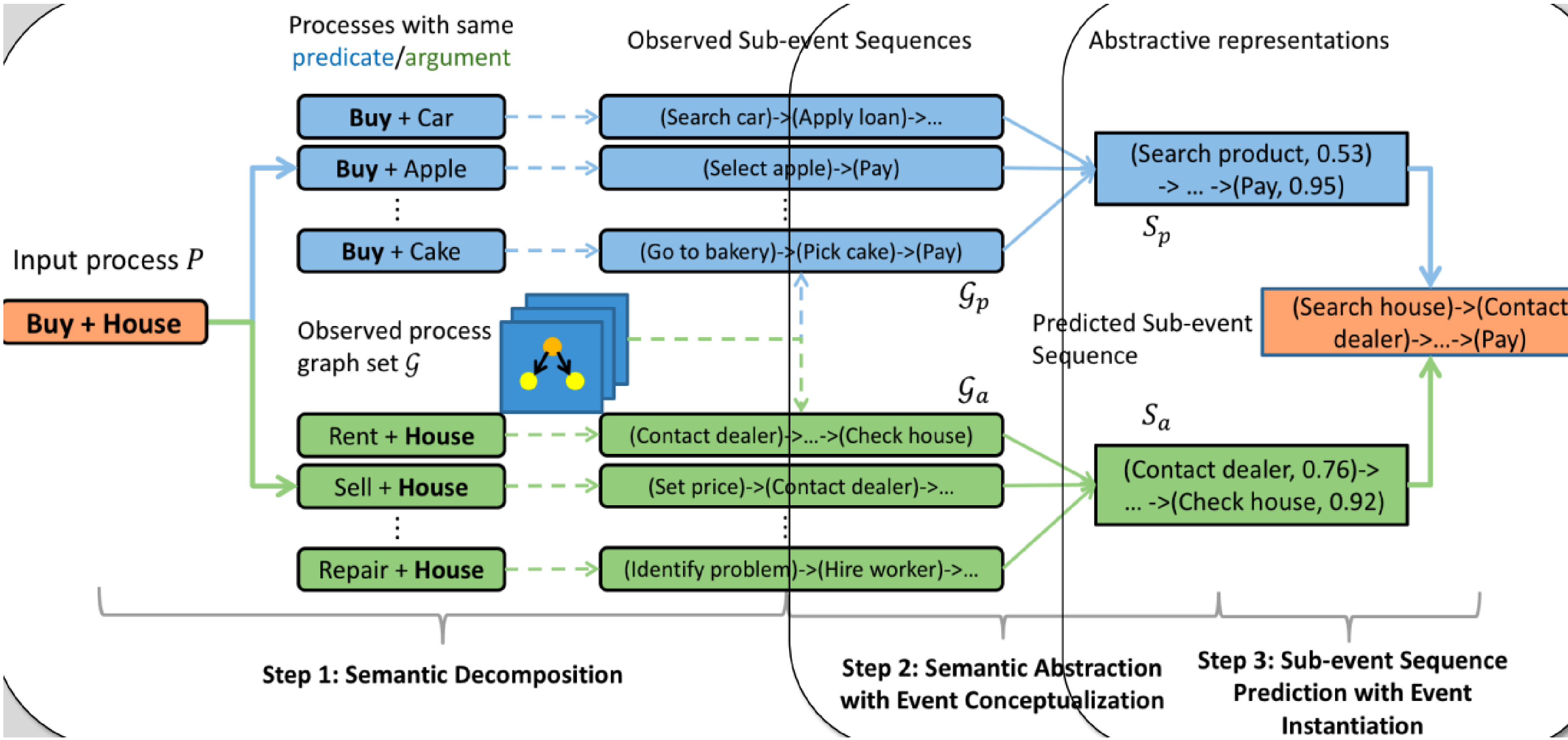
The likelihood of cholera rising is predicted high after a drought followed by storms in Angola (*based on corpus statistics*).

Predicting Events

Zhang, et al. Analogous Process Structure Induction for Sub-event Sequence Prediction. EMNLP, 2020

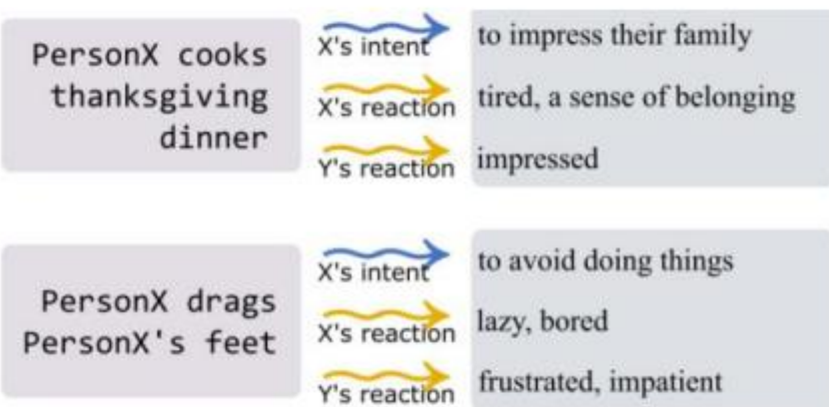


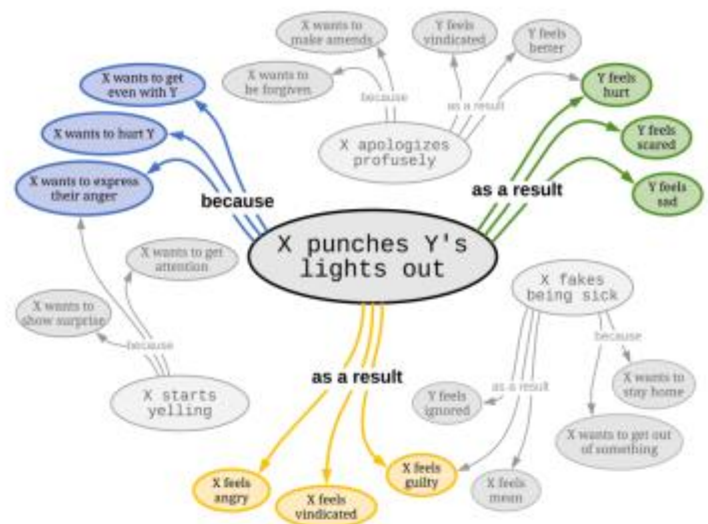
Predicting Events



Predicting Event Intention

People can easily anticipate the intents and possible reactions of participants in an event.





A commonsense-aware system should also perform such prediction.

Event2Mind – A learning system that understands stereotypical intents and reactions to events (Rashkin et al. ACL-18)

Predicting Event Intention

Is developed based on large crowdsourced corpora:

- 25,000 events
- Free-form descriptions of their intents and reactions

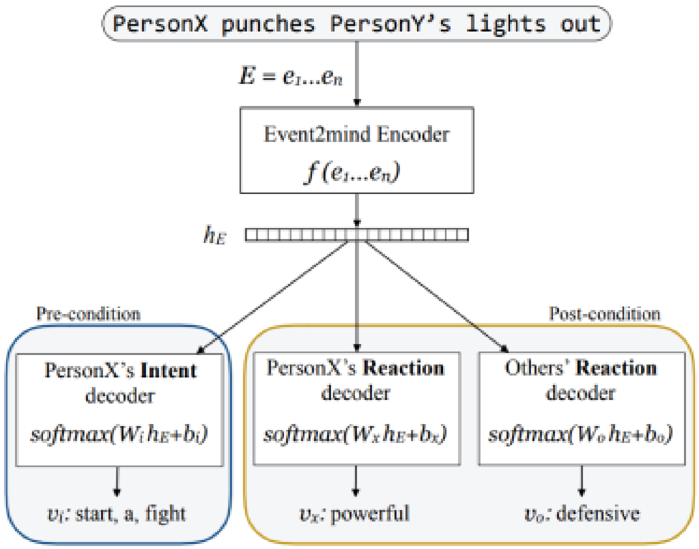
Performs Seq2NGram generation:

`PersonX cooks steak`

PersonX's intent: ["steak", "to kill their hunger", "to make dinner for the family", "to eat steak"]

PersonX's reaction: ["excited", "accomplished", "proud", "full"]

Other people's reaction: ["none", "happy", "person x cooked well."]



More follow-ups of Event2Mind

- ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning (Sap+ AAAI 2019)
- COMET: Commonsense Transformers for Automatic Knowledge Graph Construction (Bosselut+, ACL-19)

Cross lingual Event Extraction

English IE: Expensive but Generally Happy

Expert consumers define what to link/extract

↓

Human Annotators annotate a large amount of documents

↓

Researchers design features/rules, train supervised learning models

- *High Cost:* requires manual clean annotations for 500 documents
- *Poor Portability:* e.g., only covers 41 relation types and 33 event types
- Limited to a certain domain, genre, language, and data modality