Carnegie Mellon University
Language
Technologies
Institute

11-411/11-611 Natural Language Processing

Entity, Relation, and Event Extraction

David R. Mortensen and Lori Levin

April 6, 2023

Language Technologies Institute

Learning Objectives

By the end of this lecture, students will be able to...

- Describe the following tasks and resources:
  - Semantic role labeling
  - Named entity recognition
  - Relation extraction
  - Time extraction
  - Event extraction
- Describe a rule-based and/or ML-based algorithm for some of these tasks
- Describe appropriate evaluation metrics for some of these tasks

Semantic Roles

Five Sentences with Semantic Roles Held Constant

- David tossed the exams from *Pausch Bridge* to THE HILLSIDE BELOW.
- The exams were tossed by David from *Pausch Bridge* to THE HILLSIDE BELOW.
- It was David who tossed the exams from *Pausch Bridge* to THE HILLSIDE BELOW.
- The exams were thrown from *Pausch Bridge*.
- The exams were thrown to THE HILLSIDE BELOW.

Traditional Semantic Roles

- In the linguistics literature, one sees a number of common terms for semantic roles
  - Agent
  - Patient
  - Theme
  - Force
  - Experiencer
  - Stimulus
  - Recipient
  - Source
  - Goal
  - etc.

  These have their place, and are useful to know if you want to understand what a semantic role is, but are not widely used in NLP In NLP, we tend to use finer-grained (and sometimes cryptically named) semantic role labels

Traditional Semantic Roles: Examples

David tossed **the exams** from *Pausch Bridge* to THE HILLSIDE BELOW.

David	AGENT
the exams	PATIENT
Pausch Bridge	SOURCE
THE HILLSIDE BELOW	GOAL

Other types of semantic role labels

- Finegrained: writer, thing-written, eater, thing-eaten, thrower, thing-thrown, ...
- General:
  - arg0: the most agent-like argument
  - arg1: the most patient-like argument
  - argn, for other n: use as needed

One Way of Capturing the Roles of Arguments in Events Is with Semantic Role Labeling

Semantic Role Labeling: The Task

Input: a sentence, paragraph, or document.

Output: for each predicate*, labeled spans identifying each of its arguments and their roles.

*Predicates are sometimes identified in the input, sometimes not.

Propbank: an SRL Resource

- Corpus (PTB) with propositions annotated
  - Predicates (verbs)
  - Arguments (semantic roles)
- Semantic roles are Arg0, Arg1, etc., each with a description
  - Arg0 is typically the most agent-like argument
  - Labels for other arguments are somewhat arbitrary

“Agree” in PropBank

- arg0: agreer
- arg1: proposition
- arg2: other entity agreeing

- The group agreed it wouldn’t make an offer.
- Usually John agrees with Mary on everything.

“Fall (move downward)” in PropBank

- arg1: logical subject, patient, thing falling
- arg2: extent, amount fallen
- arg3: starting point
- arg4: ending point
- argM-loc: medium

- Sales fell to $251.2 million from $278.8 million.
- The average junk bond fell by 4.2%.
- The meteor fell through the atmosphere, crashing into Cambridge.

FrameNet: another SRL Resource

- A **semantic frame** is a schematic representation of a situation involving various participants, and other conceptual roles
- In **FrameNet**, frames—not verbs—are first-class citizens
  - To a first approximation, verbs that relate to the same situation belong to the same frame
  - Roles are given fine-grained labels that are specific to the frame, but not the verb
  - Frames can center around words other than verbs

The Frame change_position_on_a_scale

Core roles	
ATTRIBUTE	scalar property that the ITEM possesses
DIFFERENCE	distance by which an ITEM changes its position
FINAL_STATE	ITEM's state after the change
FINAL_VALUE	position on the scale where ITEM ends up
INITIAL_STATE	ITEM's state before the change
INITIAL_VALUE	position on the scale from which the ITEM moves
ITEM	entity that has a position on the scale
VALUE_RANGE	portion of the scale along which values of ATTRIBUTE fluctuate
Some non-core roles ...	
DURATION	length of time over which the change occurs
SPEED	rate of change of the value
GROUP	the group in which an ITEM changes the value of an ATTRIBUTE

“Triggers” for the change_position Frame

- **Verbs**: advance, climb, decline, decrease, diminish, dip, double, drop, dwindle, edge, explode, fall, fluctuate, gain, grow, increase, jump, move, mushroom, plummet, reach, rise, rocket, shift, skyrocket, slide, soar, swell, swing, triple, tumble
- **Nouns**: decline, decrease, escalation, explosion, fall, fluctuation, gain, growth, hike, increase, rise, shift, tumble
- **Adverb**: increasingly

Examples from FrameNet: change position on a scale

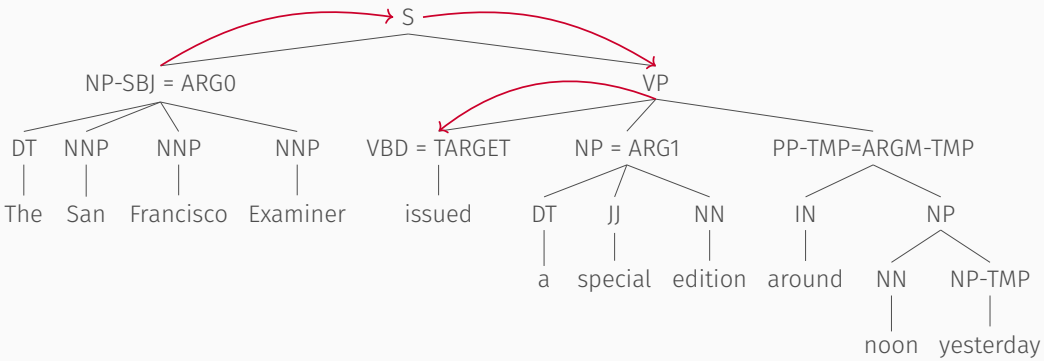
- Item: I fear **this service** will DIMINISH in quality.
- Initial and final values: Microsoft shares FELL **from 12 3/8 to 7 5/8**.
- Initial state: Diesels have INCREASED **from having a 20% market share in 1995 to just over 30%** in 2004.
- Attribute: Oil ROSE **in price** by 2%.
- Difference: Oil ROSE in price **by 2%**.
- Correlated variable: The amount of power INCREASES **with the frequency of the laser**.

An Old-Timey Approach to SRL

In some sense, SRL is a classification task.

- **Classes**: roles defined by the event type
- **Things to be classified**: the entities (encoded as noun phrases) participating in the event
- **Features**: linguistic properties of the noun phrases, path features based on the relative position of the predicate/verb and the NP, etc.

Path Features Represent the Path from Predicate to Target



Other Features used for semantic role labeling

For a phrase that you want to label, e.g., *The San Francisco Examiner*

- Phrase type: *The San Francisco Examiner* is a noun phrase
- The predicate: *issued*
- The head of the phrase: *Examiner*
- Part of speech of the head: NNP
- The voice: *issued* is in the active voice
- The position: *The San Francisco Examiner* is before *issued*
- How many arguments does the predicate need? *Issued* needs a subject and an object
- Named Entity Type: *The San Francisco Examiner* is an ORG (organization)
- The first and last words: *The* and *Examiner*

SRL and the Modern Day

In the last five years, it has been shown that traditional, linguistically informed, SRL models are not competitive with modern, end-to-end neural models (He, et al. 2017). Modern approaches treat SRL as more of a sequence labeling problem and can leverage the architectures that have been developed for such tasks.

However, other research (as by CMU's Emma Strubell) has shown that newer syntactically-aware SRL systems may be more efficient and may generalize better than the end-to-end models that are considered SOTA.

Information Extraction at a High Level

Information Extraction is one of the top ten NLP tasks in the corporate world

Turning unstructured data into structured data.

A Motivating Example

Text can be full of information, but lack the sort of structure computer programs usually use to process it. Take the following passage:

> *Citing high fuel prices, United Airlines said Friday it has 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 matched the move, spokesman Tim Wagner said. United, a unit of UAL Corp., said the increase took effect Thursday and applies to most routes where it competes against discount carriers, such as Chicago, Dallas, Denver, and San Francisco.*

We have to express this information in a **meaning representation language** like those we discussed in the previous lecture. But how do we get there?

Extracting Entities

Citing high fuel prices, [United Airlines] said Friday it has 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 matched the move, spokesman [Tim Wagner] said. [United], a unit of [UAL Corp.], said the increase took effect [Thursday] and applies to most routes where it competes against discount carriers, such as [Chicago], [Dallas], [Denver], and [San Francisco].

Extracting Relations

Citing high fuel prices, [United Airlines] said Friday it has 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 matched the move, [SPOKESMAN] [Tim Wagner] said. [United], [A UNIT OF] [UAL Corp.], said the increase took effect [Thursday] and applies to most routes where it competes against discount carriers, such as [Chicago], [Dallas], [Denver], and [San Francisco].

Extracting Events

[CITING] high fuel prices, **[United Airlines]** [SAID] Friday it has [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 [MATCHED] [THE MOVE], [SPOKESMAN] **[Tim Wagner]** [SAID]. **[United]**, [A UNIT OF] **[UAL Corp.]**, [SAID] the [INCREASE TOOK EFFECT] **[Thursday]** and [APPLIES] to most routes where it competes against discount carriers, such as **[Chicago]**, **[Dallas]**, **[Denver]**, and **[San Francisco]**.

Finding Coreference between Event Mentions

Citing high fuel prices, United Airlines said Friday it has [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 matched the [MOVE], spokesman Tim Wagner said. United, a unit of UAL Corp., said the [INCREASE] took effect Thursday and applies to most routes where it competes against discount carriers, such as Chicago, Dallas, Denver, and San Francisco.

Through tasks like these, we can convert text into meaning representation languages that can be stored in a relational or graph database and used for practical tasks

Model Based View of the Fare Increase Passage

Domain	$\mathcal{D} = \{a, b, c, d, e, f, g, h, i\}$
United, UAL, American Airlines, AMR	a, b, c, d
Tim Wagner	e
Chicago, Dallas, Denver, and San Francisco	f, g, h, i
Classes	
United, UAL, American, and AMR are organizations	$Org = \{a, b, c, d\}$
Tim Wagner is a person	$Pers = \{e\}$
Chicago, Dallas, Denver, and San Francisco are places	$Loc = \{f, g, h, i\}$
Relations	
United is a unit of UAL American is a unit of AMR	$PartOf = \{\langle a, b\rangle, \langle c, d\rangle\}$
Tim Wagner works for American Airlines	$OrgAff = \{\langle c, e\rangle\}$
United serves Chicago, Dallas, Denver, and San Francisco	$Serves = \{\langle a, f\rangle, \langle a, g\rangle, \langle a, h\rangle, \langle a, i\rangle\}$

“Experts” on the internet say that at 80% to 90% of data is unstructured. This includes medical and other important data.

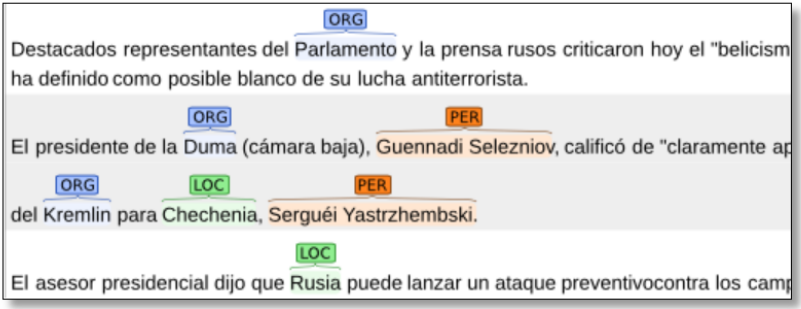
Entity Extraction: the Task

Named Entity Recognition

We have already seen the NER task

input Tokenized text

output Labeled span for each named entity in the input text



The Task of Named Entity Recognition

Elizabeth Warren, the liberal firebrand who emerged as a top Democratic contender for the **White House** on the strength of an anti-corruption platform backed by a dizzying array of policy proposals, ended her campaign on **Thursday**. A former bankruptcy law professor who forged a national reputation as a scourge of **Wall Street** even before entering politics, **Warren** had banked on a strong showing on **Super Tuesday** after a string of disappointing finishes in the early states. But she trailed far behind front-runners **Bernie Sanders** and **Joe Biden**, placing third in her home state of **Massachusetts**, which she continues to represent in the **U.S. Senate**.

...	...
B-PER	President
I-PER	Donald
I-PER	Trump
O	met
O	with
O	local
O	leaders
O	and
O	federal
O	responders
O	shortly
O	after
O	landing
O	at
O	an
B-ORG	Air
I-ORG	Force
O	base
O	in
B-LOC	Carolina
O	,
B-LOC	Puerto
I-LOC	Rico
...	...

Entity Linking

Named Entities that refer to more than one thing in the real world

- Pittsburgh, PA
- Pittsburgh, CA
- Pittsburgh neighborhood, Atlanta, GA

Knowledge Bases of Entities

There must be a unique identifier of each entity.

- Gazetteers: structured collections Geopolitical Entities (GPE) from countries to voting districts.
  - A Geopolitical Entity (GPE) as a population, border, and governemnt.
  - Locations like North Africa and the Mediterranean Region are not GPEs.
- Knowledge Graphs: graphical representations of relationships between concepts, often entities. WIKIDATA is a knowledge graph often used in entity linking.

The Entity Linking Task

inputs
1. text with named entity mentions tagged
2. a knowledge base

ouptut a mapping between entity mentions and entries in the knowledge base (unique identifiers of entities)

Never make a computer do anything you haven't tried yourself (Kevin Knight)

David and Lori have done entity linking to a database by hand in multiple languages in order to create training data in "surprise language exercises", simulations of developing NLP quickly on an emergency basis for humanitarian disasters such as floods and earthquakes.

Never make a computer do anything you haven't tried yourself (Kevin Knight)

First, you need to normalize the names of Named Entities. In languages other than English, the named entities need to undergo morphological analysis. For example, in Turkish, there are suffixes that mean "in" and "from". Place names appear in knowledge bases without the suffixes. To look up a place name, you need to take the suffix off. (Lori and David had to learn the morphology of the language in one day or less in surprise language exercises.)

Text: İstanbulda (in Istanbul)
Gazetteer: Istanbul

Then you have to find the entities to a knowledge base. The knowledge bases are very detailed, for example, covering all known geopolitical entities (GPE). (A GPE has a population, border, and government). They are based on hierarchies of geopolitical entities from countries to voting districts. First you have to decide whether something like "North Africa" is a GPE. It isn't. Then you have to associate a named entity with the right level in the hierarchy, sometimes having to disambiguate when the same name is used for a county and a city. And, of course, many cities have the same name, so you have to get the right one by seeing which state or county it is in. You have to do this for many geopolitical systems in the world.

Relation Extraction: the Task

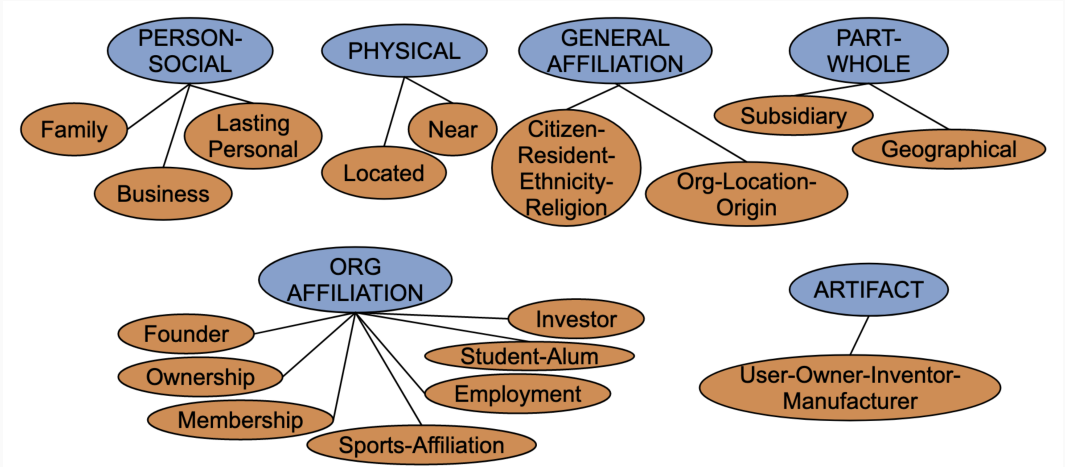
Reminder: relations are triples such as ⟨entity-1, is-employed-by, entity-2⟩.

Some databases to use as gold-standards or seeds

- DBpedia
  - Derived from Wikipedia infoboxes
  - Resource Description Framework (RDF): ⟨Golden Gate Park, location, San Francisco⟩
  - 2 Billion RDF triples
- Freebase (Wikidata): people, location, nationality
- WordNet relations: a giraffe is-a ruminant, *San Francisco* is an instance of *city*.
- TACRED: 106K hand-labeled relation triples

The ACE Relations

One historically important set of relations come from ACE (Automatic Content Extraction, a project of NIST):



Examples from TACRED

Another important relation resource is TACRED:

TACRED is a large-scale relation extraction dataset with 106,264 examples built over newswire and web text from the corpus used in the yearly TAC Knowledge Base Population (TAC KBP) challenges.

Example	Entity Types & Label
Carey will succeed Cathleen P. Black, who held the position for 15 years and will take on a new role as chairwoman of Hearst Magazines, the company said.	PERSON/TITLE Relation: *per:title*
Irene Morgan Kirkaldy, who was born and reared in Baltimore, lived on Long Island and ran a child-care center in Queens with her second husband, Stanley Kirkaldy.	PERSON/CITY Relation: *per:city_of_birth*
Baldwin declined further comment, and said JetBlue chief executive Dave Barger was unavailable.	Types: PERSON/TITLE Relation: *no_relation*

The Relation Extraction Task

inputs
1. a corpus of text
2. a set of relation types

output for each relation type, a set of tuples indicating items in that relation

Relations	Types	Examples
Physical-Located	PER-GPE	**He** was in **Tennessee**
Part-Whole-Subsidiary	ORG-ORG	**XYZ**, the parent company of **ABC**
Person-Social-Family	PER-PER	**Yoko**'s husband **John**
Org-AFF-Founder	PER-ORG	**Steve Jobs**, co-founder of **Apple**...

General Algorithm for Finding Relations

We will look at several approaches to ClassifyRelation.

function FINDRELATIONS(*words*) **returns** *relations*

relations ← *nil*
entities ← FINDENTITIES(*words*)
forall entity pairs $\langle e1, e2 \rangle$ **in** *entities* **do**
if RELATED?($e1, e2$)
relations ← *relations*+CLASSIFYRELATION($e1, e2$)

Rule-Based Relation Extraction

Hearst Rules for Relation Extraction

Hearst rules are hand-crafted relation extraction rules. The following example patterns find hyponym-hypernym relations.

NP {, NP}* {,} (and|or) other NP_H	temples, treasuries, and other important civic buildings
NP_H such as {NP,}* {(or|and)} NP	red algae such as Gelidium
such NP_H as {NP,}* {(or|and)} NP	such authors as Herrick, Goldsmith, and Shakespeare
NP_H {,} including {NP,}* {(or|and)} NP	common-law countries, including Canada and England
NP_H {,} especially {NP}* {(or|and)} NP	European countries, especially France, England, and Spain

POS and NE-Based Expressions for Relation Extraction

Modern rule-based relation extraction usually relies on more sophisticated rules that can also reference named entity types.

PER, POSITION of ORG:
George Marshall, Secretary of State of the United States

PER (named|appointed|chose|etc.) PER Prep? POSITION
Truman appointed Marshall Secretary of State

PER [be]? (named|appointed|etc.) Prep? ORG POSITION
George Marshall was named US Secretary of State

Pros and Cons of Rule-Based Information Extraction

- **Pro:** high precision
- **Con:** low recall
- **Con:** requires a lot of time from a human linguist

Feature-Based Supervised Learning of Relations

Supervised Relation Extraction

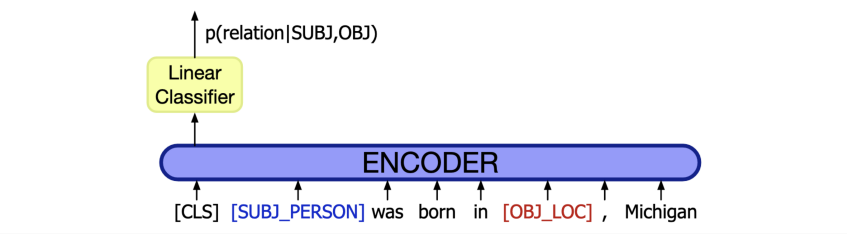
- Train on a corpus where relations are labeled
- Find pairs of text spans that are named entities
- Binary classifier: are the entities in a relation?
- If so, classify the relation

Examples of features for classifying a relation between two entities (text spans)

- The words in the two text spans
- Immediately preceding and following words
- Types of Named Entities
- Syntactic path from one entity to the other in a constituent structure tree or a dependency tree

Neural Supervised Learning

- Replace Named Entities by their NE tag (so as not to overfit to lexical items)
- Use BERT as an encoder
- Use a linear classifier to compute the probability of the relation given the entities



Pros and Cons of Supervised Relation Extraction

- **Pro:** Good when training and test data are similar
- **Con:** Bad when the training and test data are different genres such as news and social media

Bootstrapping Approaches to Relation Extraction

Bootstraping Relation Extraction: the Algorithm

function BOOTSTRAP(*Relation R*) **returns** *new relation tuples*

tuples ← Gather a set of seed tuples that have relation *R*
iterate
sentences ← find sentences that contain entities in *tuples*
patterns ← generalize the context between and around entities in *sentences*
newpairs ← use *patterns* to grep for more tuples
newpairs ← *newpairs* with high confidence
tuples ← *tuples* + *newpairs*
return tuples

Example of Bootstrapping

Establish the relation between Ryanair and Charleroi using a known pattern.

- **Sentence:** Ryanair has a hub in Charleroi.
- **Seed Relation:** <has-hub-in, Ryanair, Charleroi>
- **Seed Pattern:** `[ORG] has-a-hub-in [LOC]`

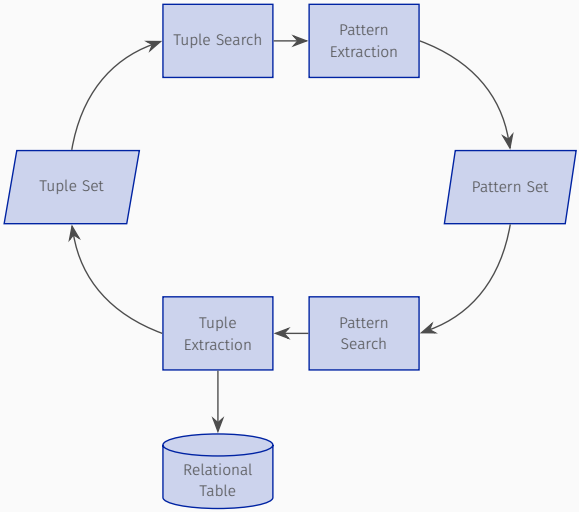
Example of Bootstrapping

Find sentences that contain Ryanair and Charleroi, assuming that any sentence containing both Ryanair and Charleroi might be saying that Ryanair has a hub in Charleroi. Make new patterns.

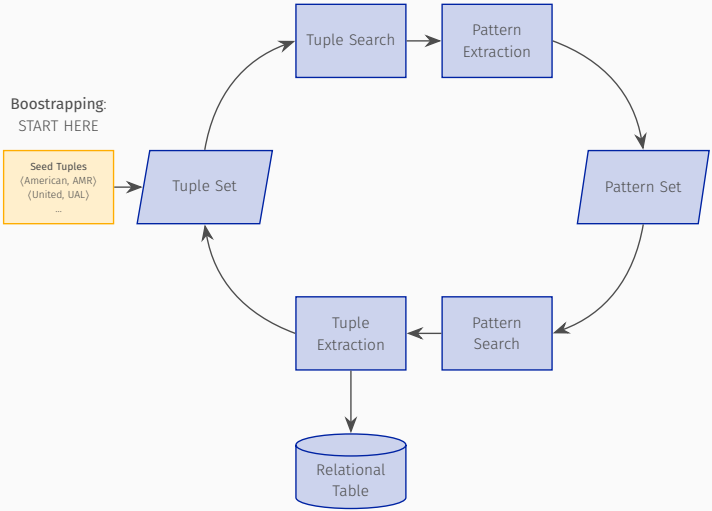
- Budget airline **Ryanair, which uses Charleroi as a hub**, scrapped all weekend flights out of the airport.
  `[ORG] which uses [LOC] as a hub`
- All flights in and out of **Ryanair's hub at Charleroi airport** were grounded on Friday.
  `[ORG]'s hub at [LOC]`
- A spokesman at **Chareleroi, a main hub for Ryanair**, estimated that 8000 passengers had already been affected.
  `[LOC], a main hub for [ORG]`

Assign confidence scores to the new patterns.
Use the new patterns to find additional has-hub-in tuples for other airline-city pairs.

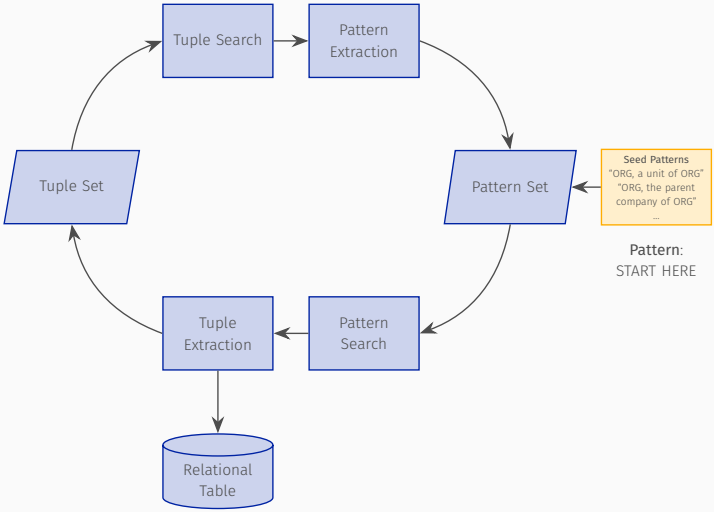
Cycle for Bootstrapping-based and Pattern-based Relation Extraction



Cycle for Bootstrapping-based and Pattern-based Relation Extraction



Cycle for Bootstrapping-based and Pattern-based Relation Extraction



Pros and Cons of Using Seeds

- **Pro:** you can start with very little labelled data
- **Con:** Drift: a wrong pattern can lead to a cascade of wrong learning
  - Consider the entities **America** and **AMR** (AMR Corp. or Abstract Meaning Representation)
  - **American** is-subsidiary **AMR**
  - Take the sentence “the loyal **American** used **AMR** to encode semantics”
  - The model assumes that “X used Y” means “X is-subsidiary Y,” yielding both spurious tuples and spurious patterns
  - This is why low-confidence patterns are ignored

Contemporary Relation Extraction

Distant Supervision for Relation Extraction: the Algorithm

function DISTANT SUPERVISION(*Database D, Text T*) **returns** *relation classifier C*

foreach relation *R*
  foreach tuple (*e1,e2*) of entities with relation *R* in *D*
    sentences ← Sentences in *T* that contain *e1* and *e2*
    f ← Frequent features in *sentences*
    observations ← observations + new training tuple (*e1, e2, f, R*)
C ← Train supervised classifier on *observations*
return *C*

Distant Supervision: use existing databases as big seeds

- Start with no labelled data
- Get tuples out of Freebase or DBPedia:
  ⟨`date-of-birth, Albert Einstein, Ulm`⟩
- Run a named entity tagger on a lot of text
- Extract all sentences that contain Albert Einstein and Ulm
- Extract frequent features from those sentences
- Use tuples of the entities, the features, and the relation as training instances

Pros and Cons of Distant Supervision

- **Con:** low precision
- **Pro:** since it doesn't use a training corpus, it isn't as sensitive to genre (e.g., news vs social media) as fully supervised models
- **Pro:** create enough data to train neural classifiers, which do not need feature extraction

Evaluation of Relation Extraction

Supervised Relation Extraction is Evaluated with Standard Tools

- Human-annotated, gold-standard relations
- Metrics
  - Precision
  - Recall
  - F-measure
- Labeled versus unlabeled precision and recall

  labeled must predict existence of relation and type of relation

  unlabeled must predict existence of relation only

Unsupervised and Semi-Supervised Relation Extraction is Difficult to Evaluate

- Use very large amounts of text (no small, labeled test set or pre-annotated instances)
- Solution: sampling outputs for human evaluation
  - Based on TUPLES not MENTIONS (what is in the knowledge base when all is said and done)
  - Estimated precision:

$$\hat{P} = \frac{\text{\# of correctly extracted relation tuples in sample}}{\text{total \# of extracted relation tuples in the sample}} \quad (1)$$

Extracting Times

The Time Extraction Task

input text

output a normalized representation of the time expressions from the text

Examples of Temporal Expressions

Absolute	Relative	Durations
April 24, 1916	yesterday	four hours
The summer of ’77	next semester	three weeks
10:15 AM	two weeks from yesterday	six days
The 3rd quarter of 2006	last quarter	the last three quarters

Normalizing Time Expressions

Most temporal relations are underspecified.

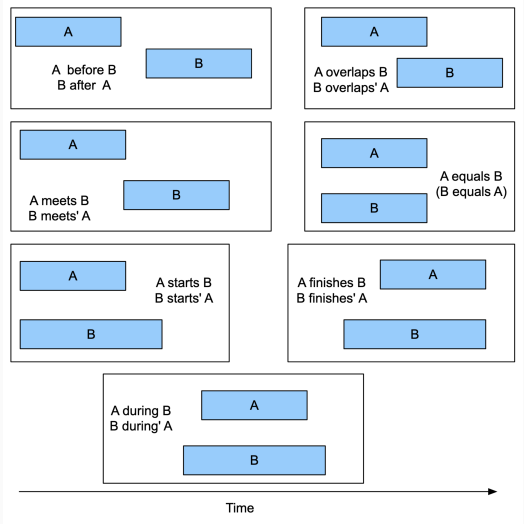
`Conversation 1:`
A: I can't make our lunch meeting on Monday next week.
B: Ok. Let's meet on Tuesday.

`Conversation 2:`
(Background: the conference usually takes place on the second Monday in February.)
A: Next year the conference coincides with the lunar new year.
B: Ok. So we have to postpone by a week, but that is Presidents Day, so let's meet on Tuesday.

What is the time of the meeting, the duration of the meeting, and the exact date, including the year, of the meeting?

The Allen Temporal Relations

The Allen Temporal Relations are a widely used scheme for characterizing the time relationship between two events.



TimeBank: an annotated corpus of temporal expressions

- 183 news articles, overlapping with PennTreebank and PropBank
- TimeML (Time Markup Language):

Example from TimeBank and Allen Relations

10/26/89
Delta Air Lines earnings soared 33% to a record in the fiscal first quarter, bucking the industry trend toward declining profits.

How do you know that the earning event is during the first quarter?
How do you know that the earning event is before the date of the article?
How do you know that earning and bucking are at the same time?
How do you know that the earning event is during the declining event?

```
<TIMEX3 tid="t57" type="DATE" value="1989-10-26"
functionInDocument="CREATION_TIME">
10/26/89 </TIMEX3>

Delta Air Lines earnings
<EVENT eid="e1" class="OCCURRENCE"> soared </EVENT>
33% to a record in
<TIMEX3 tid="t58" type="DATE" value="1989-Q1" anchorTimeID="t57">
fiscal first quarter </TIMEX3>,
<EVENT eid="e3" class="OCCURRENCE">bucking</EVENT>
the industry trend toward
<EVENT eid="e4" class="OCCURRENCE">declining</EVENT>
profits.
```

Also, e1 is included in t58. E1 is before t57. e1 is simultaneous with e3. e4 includes e1.

Triggers for Time Expressions

Time extraction models, whether rule- or ML-based, rely upon (or learn) certain triggers for times and temporal relations.

Category	Examples
Noun	*morning*, *noon*, *night*, *winter*, *dusk*, *dawn*
Proper Noun	*January, Monday, Ides, Easter, Rosh Hashana, Ramadan, Tet*
Adjective	*recent, past, annual, former*
Adverb	*hourly, daily, monthly, yearly*

Event Extraction: the Task

The Event Extraction Task

input text

output a structured representation of the events in the text, their types, their arguments and (often) the roles of the arguments

An Example Event Template

FARE-RAISE ATTEMPT:	LEAD AIRLINE:	UNITED AIRLINES
	AMOUNT:	$6
	EFFECTIVE DATE:	2006-10-26
	FOLLOWER:	AMERICAN AIRLINES

Early Approaches to Event Extraction

One Early Approach to Event Detection Uses Finite-State Transducers

The FASTUS system used a cascade* of FSTs (which we told you you'd see again) to extract ever-more abstract units of language.

- Tokenization
- Multiword expressions
- Shallow syntax (FSTs can be used for that too!)
- Shallow semantics and event coreference

*A cascade, as the word is used here, is a sequence of composed FSTs.

No.	Step	Description
1	**Tokens**	Tokenize input stream of characters
2	**Complex Words**	Multiword phrases, numbers, and proper names.
3	**Basic phrases**	Segment sentences into noun and verb groups
4	**Complex phrases**	Identify complex noun groups and verb groups
5	**Semantic Patterns**	Identify entities and events, insert into templates.
6	**Merging**	Merge references to the same entity or event

Pros and Cons of FASTUS

Pro: Computationally efficient

Pro: Interpretable and relatively easy to unit test and debug

Con: Extensive use of handcrafted resources and transducers

Con: Syntactic generalization is very shallow

Machine Learning Approaches to Event Extraction

Features for Event Detection

Most event detection systems, whether rule- or ML-based, rely on a set of morphological, syntactic, and semantic features:

Feature	Explanation
Character affixes	Character-level prefixes and suffixes of target word
Nominalization suffix	Character-level suffixes for nominalizations (e.g., *-tion*)
Part of speech	Part of speech of the target word
Light verb	Binary feature indicating that the target is governed by a light verb
Subject syntactic category	Syntactic category of the subject of the sentence
Morphological stem	Stemmed version of the target word
Verb root	Root form of the verb basis for a nominalization
WordNet hypernyms	Hypernym set for the target

Neural systems may also use embeddings.

ML-Based Template Filling Event Extraction

inputs
- training documents with text spans annotated with predefined templates + slot fillers
- unlabeled documents

output set of templates, one per each input event, with slots filled by text spans

This is often achieved by combining two systems:

1. TEMPLATE RECOGNITION
2. ROLE-FILLER EXTRACTION

Template Recognition System

- Also called EVENT RECOGNITION
- Classification task

  input features extracted from from every sequence labeled in the training set

  output template for corresponding passage

Role-Filler Extraction System

- Also a classification task

  input each noun phrase in the input document OR a sequence of words (to be labeled)

  output classes for each of the phrases/spans in the input

Semantic Role Labeling (SRL) is a special case of Role-Filler Extraction.

Skip to the End.

Information extraction is commercially important but also technically and scientifically interesting—there are many competing ideas and approaches that bring linguistics, machine learning, and other fields together.

Questions?