

# Information Extraction

LING 570

Fei Xia

# Outline

- What is IE?
- General process
- Relation detection
  - Supervised method
  - Lightly supervised method
- J&M Chapter 22

# What is IE?

- Motivation:
  - Unstructured data hard to manage, assess, analyze
  - Many tools for manipulating structured data:
    - Databases: efficient search, agglomeration
    - Statistical analysis packages
    - Decision support systems
- Goal:
  - Given unstructured information in texts
  - Convert to structured data
    - E.g., populate DBs

# An example

Citing high fuel prices, United Airlines said Friday it has increased fares by \$6 per round trip 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 to Dallas and Denver to San Francisco.

Announcement date:	10/14/2016 (Friday)
Lead Airlines:	United Airlines
Amount:	\$6 round trip
Effective date:	10/13/2016 (Thursday)
Follower:	American Airlines

# Applications

- Shared tasks:
  - Message Understanding Conferences (MUC)
    - Joint ventures/mergers:
      - CompanyA, CompanyB, CompanyC, amount, type
    - Terrorist incidents:
      - Incident type, Actor, Victim, Date, Location
  - BioNLP contests
- General domains:
  - Pricegrabber
  - Stock market analysis
  - Apartment finder

# Approaches

- Incorporates range of techniques and tools:
  - FSAs (pattern extraction)
  - Supervised learning:
    - Classification
    - Sequence labeling
  - Semi- and un-supervised learning: clustering, bootstrapping

# Common pre-processing steps

- Part-of-Speech tagging
- Named Entity Recognition
- Chunking
- Parsing
- ...

# Outline

- What is IE?
- General process
- Relation detection
  - Supervised method
  - Lightly supervised method



# The general process (1)

- Tokenization, POS tagging, chunking, parsing, ...
- NER: detect entities mentioned in a text
  - Often application-specific
- Citing high fuel prices, [ORG United Airlines] said [TIME Friday] it has increased fares by [MONEY \$6] per round trip to some cities also served by lower-cost carriers. [ORG American Airlines], a unit of [ORG AMR Corp.] immediately matched the move, spokesman [PER Time Wagner] said. [ORG United], a unit of [ORG UAL Corp.], said the increase took effect [TIME Thursday] and applies to most routes where it competes against discount carriers, such as [LOC Chicago] to [LOC Dallas] and [LOC Denver] to [LOC San Francisco].

# The general process (

- Reference resolution: “United” refers “United Airlines”
  - Create entity equivalent classes
- Citing high fuel prices, **[<sub>ORG</sub> United Airlines]** said [<sub>Time</sub> Friday] **it** has increased fares by [<sub>MONEY</sub> \$6] per round trip to some cities also served by lower-cost carriers. [<sub>ORG</sub> American Airlines], a unit of [<sub>ORG</sub> AMR Corp.] immediately matched the move, spokesman [<sub>PER</sub> Time Wagner] said. **[<sub>org</sub> United]**, a unit of [<sub>ORG</sub> UAL Corp.], said the increase took effect [<sub>TIME</sub> Thursday] and applies to most routes where **it** competes against discount carriers, such as [<sub>LOC</sub> Chicago] to [<sub>LOC</sub> Dallas] and [<sub>LOC</sub> Denver] to [<sub>LOC</sub> San Francisco].

# The general process (3)

- Relation detection and classification:
  - Find and classify relations among entities in text
    - Mix of application-specific and generic relations
    - Generic relations:
      - Family, employment, part-whole, membership, geospatial
  - Generic relations:
    - United is **part-of** UAL Corp.
    - American Airlines is **part-of** AMR
    - Tim Wagner is an **employee** of American Airlines
  - Domain-specific relations:
    - United **serves** Chicago; United **serves** Denver, etc

# The general process (4)

- Event detection and classification:
  - Find key events
  - Fare increase by United
  - Matching fare increase by American
  - Reporting of fare increases
    - How cued?
      - Instances of “said” and ”cite”

# The general process (5)

- Temporal expression recognition
  - Finds temporal events in text
    - E.g. Friday, Thursday in example text
  - Others:
    - Date expressions:
      - Absolute: days of week, holidays
      - Relative: two days from now, next year
    - Clock times:
      - noon, 3:30pm
- Temporal analysis:
  - Map expressions to points/spans on timeline
    - Anchor: e.g. Friday, Thursday: tied to example dateline

# The general process (6)

- Template filling: Texts describe stereotypical situations in domain
  - Identify texts evoking template situations
  - Fill slots in templates based on text
  - In example: fare-raise is stereo-typical event

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

# The general process (recap)

- NER
- Reference resolution
- Relation detection and classification
- Event detection and classification
- Temporal expression detection and temporal analysis
- Template-filling

# Outline

- What is IE?
- General process
- Relation detection
  - Supervised method
  - Lightly supervised method



# Common relations

Relations		Examples	Types
Affiliations	Personal	<i>married to, mother of</i>	$\text{PER} \rightarrow \text{PER}$
	Organizational	<i>spokesman for, president of</i>	$\text{PER} \rightarrow \text{ORG}$
	Artifactual	<i>owns, invented, produces</i>	$(\text{PER} \mid \text{ORG}) \rightarrow \text{ART}$
Geospatial	Proximity	<i>near, on outskirts</i>	$\text{LOC} \rightarrow \text{LOC}$
	Directional	<i>southeast of</i>	$\text{LOC} \rightarrow \text{LOC}$
Part-Of	Organizational	<i>a unit of, parent of</i>	$\text{ORG} \rightarrow \text{ORG}$
	Political	<i>annexed, acquired</i>	$\text{GPE} \rightarrow \text{GPE}$

# Relations and entities from the example

## Domain

United, UAL, American Airlines, AMR  
Tim Wagner  
Chicago, Dallas, Denver, and San Francisco

$\mathcal{D} = \{a, b, c, d, e, f, g, h, i\}$   
 $a, b, c, d$   
 $e$   
 $f, g, h, i$

## Classes

United, UAL, American, and AMR are organizations  
Tim Wagner is a person  
Chicago, Dallas, Denver, and San Francisco are places

$Org = \{a, b, c, d\}$   
 $Pers = \{e\}$   
 $Loc = \{f, g, h, i\}$

## Relations

United is a unit of UAL  
American is a unit of AMR  
Tim Wagner works for American Airlines  
United serves Chicago, Dallas, Denver, and San Francisco

$PartOf = \{\langle a, b \rangle, \langle c, d \rangle\}$   
 $OrgAff = \{\langle c, e \rangle\}$   
 $Serves = \{\langle a, f \rangle, \langle a, g \rangle, \langle a, h \rangle, \langle a, i \rangle\}$

# Supervised Approaches to Relation Analysis

- Two-stage process:
  - Relation detection:
    - Detect whether a relation holds between two entities
      - Positive examples:
        - » Drawn from labeled training data
      - Negative examples:
        - » Generated from within-sentence entity pairs with no relation
  - Relation classification:
    - Label each related pair of entities by type
      - Multi-class classification task

# Some Example Features

[<sub>ORG</sub> American Airlines], a unit of [<sub>ORG</sub> AMR Corp.] immediately matched the move, spokesman [<sub>PER</sub> Time Wagner] said.

## Entity-based features

Entity <sub>1</sub> type	ORG
Entity <sub>1</sub> head	<i>airlines</i>
Entity <sub>2</sub> type	PERS
Entity <sub>2</sub> head	<i>Wagner</i>
Concatenated types	ORGPERS

## Word-based features

Between-entity bag of words	{ <i>a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman</i> }
Word(s) before Entity <sub>1</sub>	NONE
Word(s) after Entity <sub>2</sub>	<i>said</i>

## 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	<i>Airlines</i> $\leftarrow_{subj}$ <i>matched</i> $\leftarrow_{comp}$ <i>said</i> $\rightarrow_{subj}$ <i>Wagner</i>

# Lightly Supervised Approaches

- Supervised approaches is highly effective, **but**
  - require large manually annotated corpus for training
- Unsupervised approaches interesting, **but**
  - may not yield results consistent with desired categories
- Lightly supervised approaches:
  - Build on small seed sets of patterns that capture relations of interest
    - E.g. manually constructed regular expressions
  - Iteratively generalize and refine patterns to optimize

# Bootstrapping Relations

- Create seed patterns and instances
  - Example: finding airline hub cities
    - Initial pattern: / \* has a hub at \* /
  - In a large corpus finds examples like:
    - Delta has a hub at LaGuardia.
    - Milwaukee-based Midwest has a hub at KCI.
    - American airlines has a hub at the San Juan airport
  - Instances:
    - (Delta, LaGuardia), (Midwest, KCI), (American, San Juan)

# Patterns → more tuples

ORG has a hub at LOC

Milwaukee-based Midwest has a hub at KCI.

Delta has a hub at LaGuardia.

Bulgaria Air has a hub at Sofia Airport, as does Hemus Air.

American Airlines has a hub at the San Juan airport.

No frills rival easyJet, which has established a hub at Liverpool...

Ryanair also has a continental hub at Charleroi airport (Belgium).

# Tuples → more patterns

(hub, Ryanair, Charleroi)

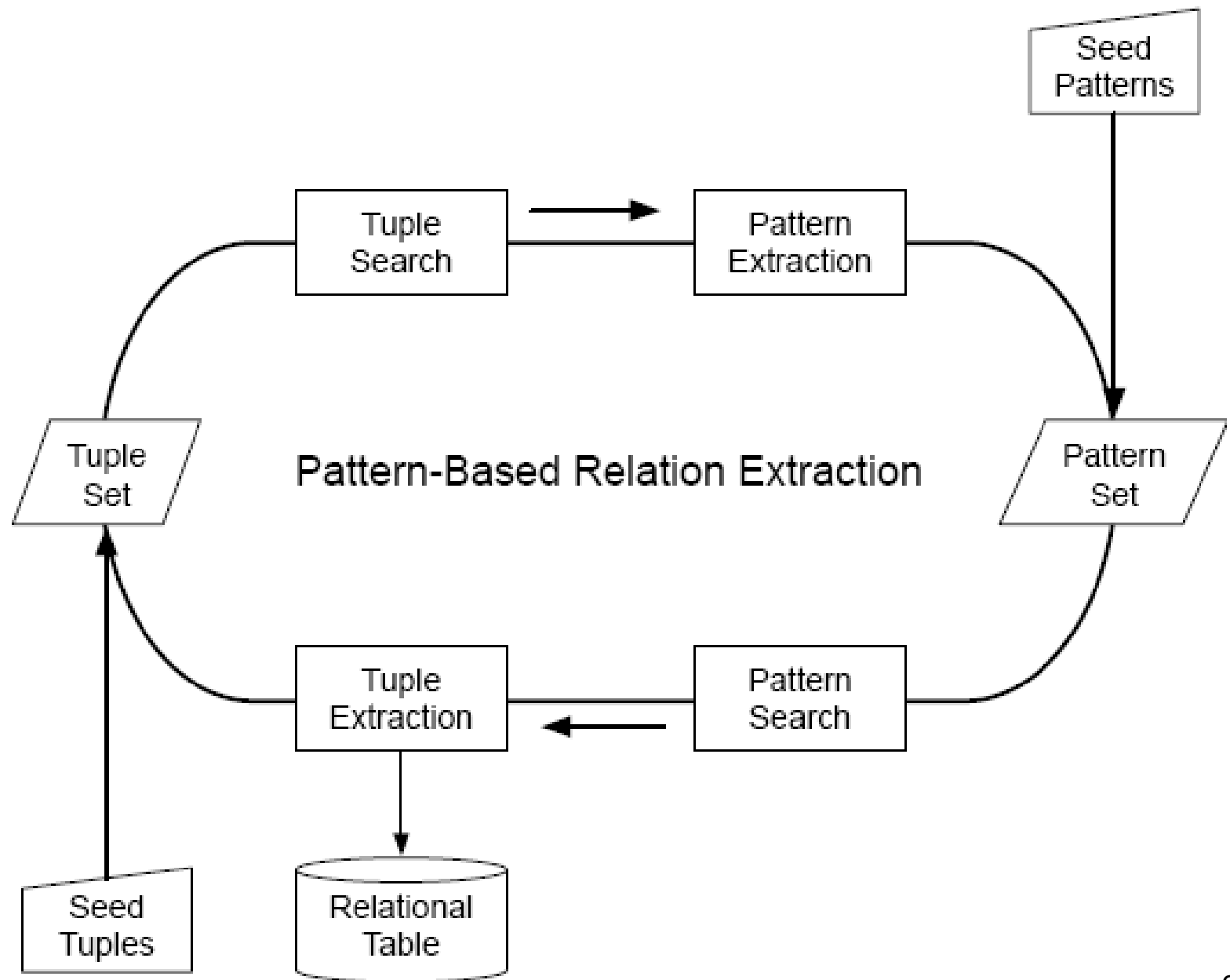
Budget airline Ryanair, which uses Charleroi as a hub, scrapped all weekend flights out of the airport.

All flights in and out of Ryanair's Belgian hub at Charleroi airport were grounded on Friday...

A spokesman at Charleroi, a main hub for Ryanair, estimated that 8000 passengers had already been affected.

```
/ [ORG], which uses [LOC] as a hub /  
/ [ORG]'s hub at [LOC] /  
/ [LOC] a main hub for [ORG] /
```





# Bootstrapping Relations

- Issues:
  - False alarms:
    - Some matches aren't valid relations
      - The catheter has a hub at the very end of the hall.
    - Fix? Change the pattern to /[ORG] has hub at [LOC]/
  - Misses:
    - Some instances are narrowly missed by pattern
      - No frills rival easyJet, which has **established** a hub at Liverpool.
      - Ryanair also has a **continental** hub at Charleroi airport.
    - Fix?
      - Relax patterns – see the 1<sup>st</sup> issue
      - Add new high-precision patterns

# Bootstrapping

- How can we obtain more high quality patterns?
  - Human labeling? Expensive
  - Use tuples identified by seed patterns
  - Find other occurrences of tuples
  - Extract patterns from those occurrences
- E.g. tuple: (hub, Ryanair, Detroit)
  - Occurrences:
    - Budget airline Ryanair, which uses Detroit as a hub
    - All flights in and out of Ryanair's Belgian hub at Detroit
  - Patterns:
    - /[ORG], which uses [LOC] as a hub /
    - /[ORG]'s hub at [LOC] /

# Relation Analysis Evaluation

- Approaches:
  - Standardized, labeled data sets
    - Precision, recall, f-measure
      - ‘Labeled’ precision: test both detection and classification
      - ‘Unlabeled’ precision: test only detection
  - No fully labeled corpus
    - Evaluate retrieved tuples (e.g. DB contents)
      - Don’t care how many times we find the same tuple