Information Extraction

LING 570

Fei Xia

Week 11: 12/07/09

Outline

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

What is IE?

- The task: turn the unstructured semantic information hidden in texts into structured data.

An example

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 to Dallas and Denver to San Francisco.

Announcement date: 10/20/06 (Friday)
Lead Airlines: United Airlines
Amount: $6 round trip
Effective date: 10/19/06 (Thursday)
Follower: American Airlines

Other examples

- Apartment finder
- Job finder
- Stock analysis
- …

The general process

- NER: detect and classify all the proper names mentioned in a text.
- Reference resolution: “United” refers “United Airlines”
- Relation detection and classification: to find and classify semantic relations among the entities discovered in a given text:
  - Ex: “United” is a part of “UAL”, “Tim Wagner” is an employee of AA.

The general process (cont)

- Event detection and classification
  - The fare increase by United
  - The fare match by AA
  - The two uses of “said” and the use of “cite”
- Temporal expression detection and temporal analysis: Friday, Tuesday, …
- Template-filling

The filled templates

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

Relation detection

Typical semantic relations

Relations		Examples	Types
Affiliations			
	Personal	*married to, mother of*	PER → PER
	Organizational	*spokesman for, president of*	PER → ORG
	Artifactual	*owns, invented, produces*	(PER | ORG) → ART
Geospatial			
	Proximity	*near, on outskirts*	LOC → LOC
	Directional	*southeast of*	LOC → LOC
Part-Of			
	Organizational	*a unit of, parent of*	ORG → ORG
	Political	*annexed, acquired*	GPE → GPE

Relations in the sample text

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	$PartOf = \{\langle a,b\rangle, \langle c,d\rangle\}$
American is a unit of AMR	
Tim Wagner works for American Airlines	$OrgAff = \{\langle c,e\rangle\}$
United serves Chicago, Dallas, Denver and San Francisco	$Serves = \{\langle a,f\rangle, \langle a,g\rangle, \langle a,h\rangle, \langle a,i\rangle\}$

Supervised learning

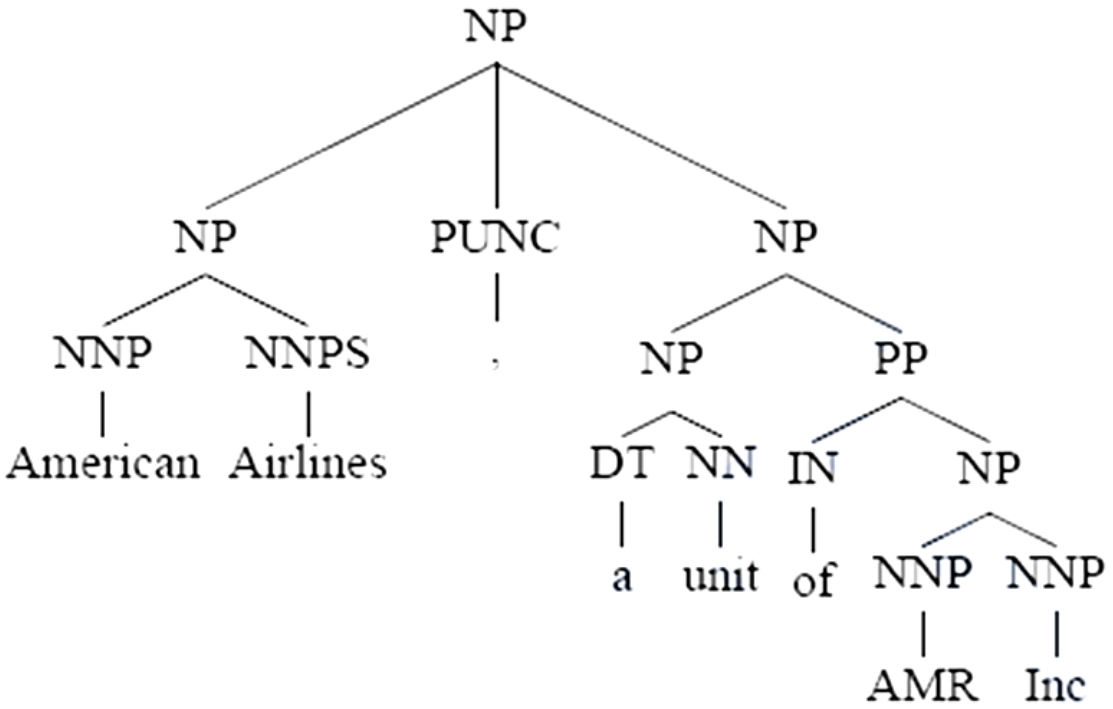
- Two-stage approach:
  - Detect whether a relation is present between two entities: a binary classification task
    - Create positive and negative examples
  - Classify any detected relations: a multi-class task

Features

- Named-entity features:
  - NE types of the two candidate arguments
  - Head words of the arguments
  - Bag of words from each of the arguments
- Features derived from the words in the text
  - The bag of words between the entities
  - Words immediately preceding/following the entities
  - Distance in words between the arguments
  - Number of entities between the arguments

Features (cont)

- Syntactic structure features:



served by lower-cost carriers. [$_{ORG}$ American Airlines], a unit of [$_{ORG}$ AMR Corp.], immediately matched the move, spokesman [$_{PERS}$ Tim Wagner] said.

Entity-based features	
Entity$_1$ type	ORG
Entity$_1$ head	*airlines*
Entity$_2$ type	PERS
Entity$_2$ head	*Wagner*
Word-based features	
Between-entity bag of words	{ *a, unit, of, AMR, Inc., spokesman* }
Word(s) before Entity$_1$	NONE
Word(s) after Entity$_2$	*said*

Syntactic features

Typed-dependency path: *Airlines* $\leftarrow_{subj}$ *matched* $\leftarrow_{comp}$ *said* $\rightarrow_{subj}$ *Wagner*

Lightly supervised approaches

- Annotating data is expensive.

- Main ideas:
  - Create seed patterns and seed tuples.
    - Seed pattern: ORG has a hub at LOC
    - Seed tuples: (hub, Northwest, Detroit)
  - Repeat
    - Apply patterns to find more tuples, and apply tuples to find more patterns
    - Assess the newly discovered patterns and tuples

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

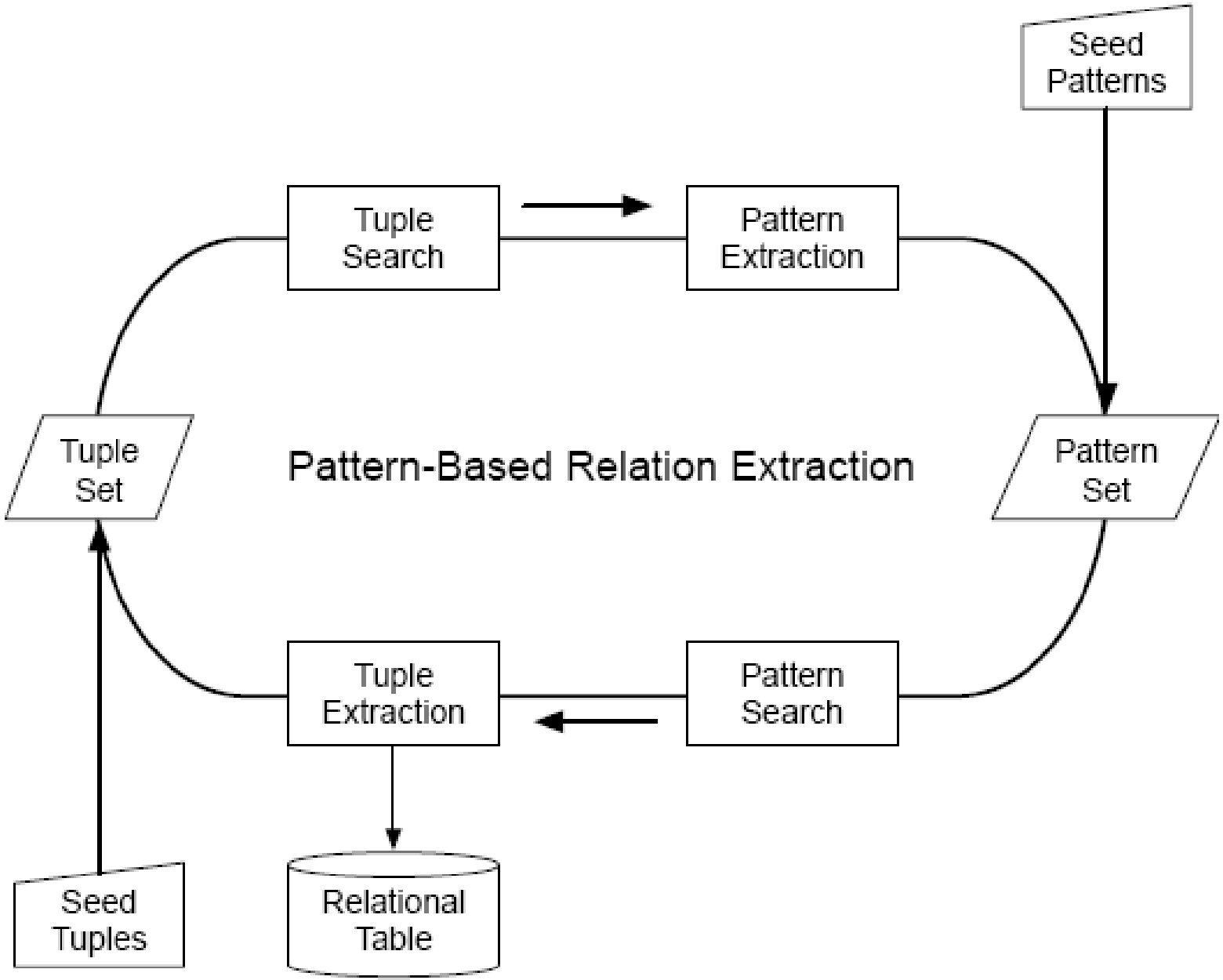
(hub, Ryanair, Charleroi)

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] /
```


Seed Patterns
Tuple Search
Pattern Extraction
Tuple Set
Pattern-Based Relation Extraction
Pattern Set
Tuple Extraction
Pattern Search
Seed Tuples
Relational Table

Outline

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

Next quarter

- LING 575: IE in the medical domain
- Input: medical clinical data
- Output: medication information:
  - medication name
  - dosage, mode, frequency, reason, etc.

The i2b2 Challenge Task

Extract medication information from discharge summaries

Record #111999

…

TREATMENT:

After observing high blood sugar , patient was given 150 cc insulin once a day for one week.

…

ALLERGIES:

Sulfa

…

DISCHARGE MEDICATIONS:

Tylenol 2 tabs q.d. p.o. headache

...