Relation Extraction

Pawan Goyal

CSE, IIT Kharagpur

Week 10, Lecture 4

Bootstrapping approaches

- If you don't have enough annotated text to train on ...
- But you do have:
  - some seed instances of the relation
  - (or some patterns that work pretty well)
  - and lots and lots of unannotated text (e.g., the web)
- can you use those seeds to do something useful?
- Bootstrapping can be considered semi-supervised

Bootstrapping example

- Target relation: burial place

Bootstrapping example

- Target relation: burial place
- Seed tuple : [*Mark Twain, Elmira*]

Bootstrapping example

- Target relation: burial place
- Seed tuple : [*Mark Twain, Elmira*]
- Google for “Mark Twain” and “Elmira”

Bootstrapping example

- Target relation: burial place
- Seed tuple : [*Mark Twain, Elmira*]
- Google for “Mark Twain” and “Elmira”

“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

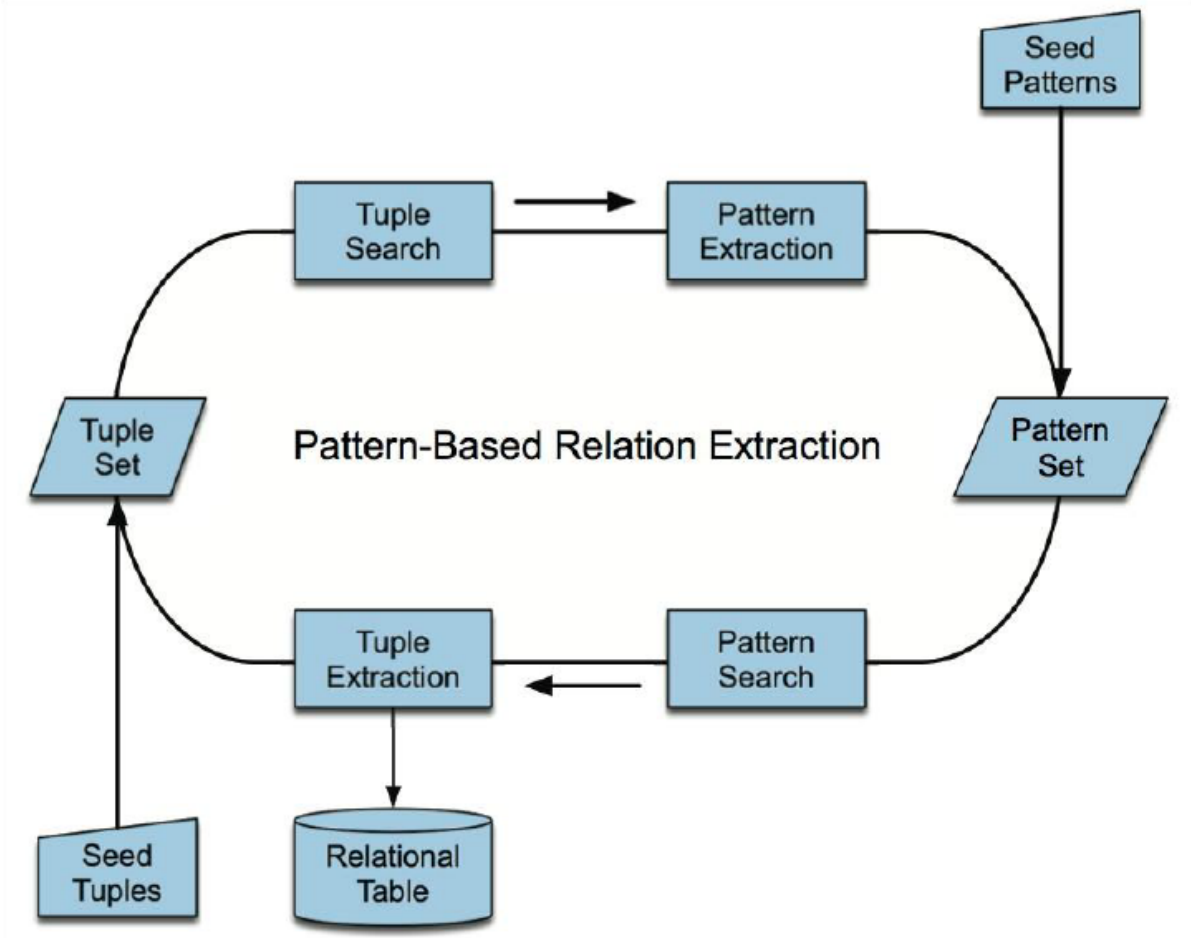
Bootstrapping example

- Target relation: burial place
- Seed tuple : [*Mark Twain, Elmira*]
- Google for “Mark Twain” and “Elmira”

“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 search for new tuples

Bootstrapping relations



Bootstrapping problems

- Requires that we have seeds for each relation
  - Sensitive to original set of seeds
- Generally have lots of parameters to be tuned
- No probabilistic interpretation
  - Hard to know how confident to be in each result

Supervised Relation Extraction

- Choose a set of relations you would like to extract

Supervised Relation Extraction

- Choose a set of relations you would like to extract
- 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

Supervised Relation Extraction: An extra step helps

- Find all pairs of named entities (usually in same sentence)

Supervised Relation Extraction: An extra step helps

- Find all pairs of named entities (usually in same sentence)
- **Extra step:** Build a binary classifier to decide if 2 entities are related

Supervised Relation Extraction: An extra step helps

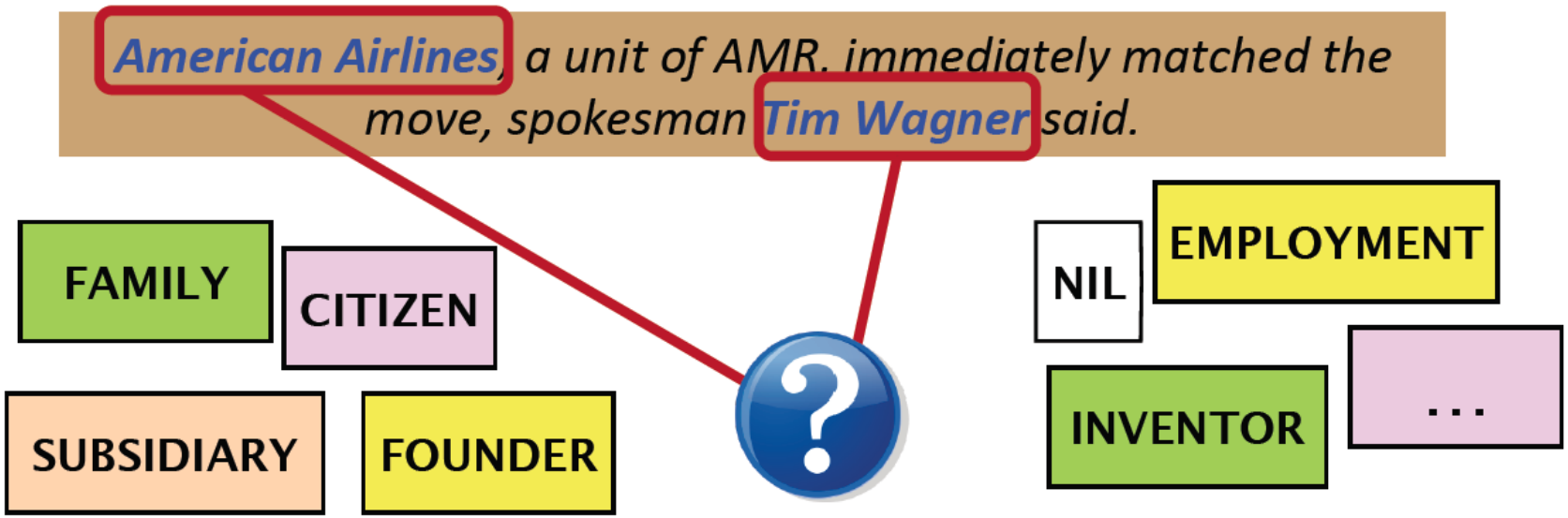
- Find all pairs of named entities (usually in same sentence)
- **Extra step:** Build a binary classifier to decide if 2 entities are related
- If yes, use another classifier to classify the relation

Why the extra step?

- Faster classification training by eliminating most pairs
- Can use distinct feature-sets appropriate for each task

Classify the relation between two entities in a sentence

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



Features: words in mentions M1 and M2

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

Features: words in mentions M1 and M2

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

Bag-of-words features

WM1 = {American, Airlines}, WM2 = {Tim, Wagner}

Features: words in mentions M1 and M2

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

Bag-of-words features

WM1 = {American, Airlines}, WM2 = {Tim, Wagner}

Head-word features

HM1 = Airlines, HM2 = Wagner, HM12 = Airlines+Wagner

Features: word around the mentions

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

Features: word around the mentions

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

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

M2:-1 = spokesman, M2: +1 = said

Features: word around the mentions

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

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}

Named Entity Type and Mention Level Features

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

Named Entity Type and Mention Level Features

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

Named-entity types

M1-NE = ORG, M2-NE = PERSON

Named Entity Type and Mention Level Features

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

Named-entity types

M1-NE = ORG, M2-NE = PERSON

Concatenation of the two named-entity types

M12-NE = ORG-PERSON

Named Entity Type and Mention Level Features

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

Named-entity types

M1-NE = ORG, M2-NE = PERSON

Concatenation of the two named-entity types

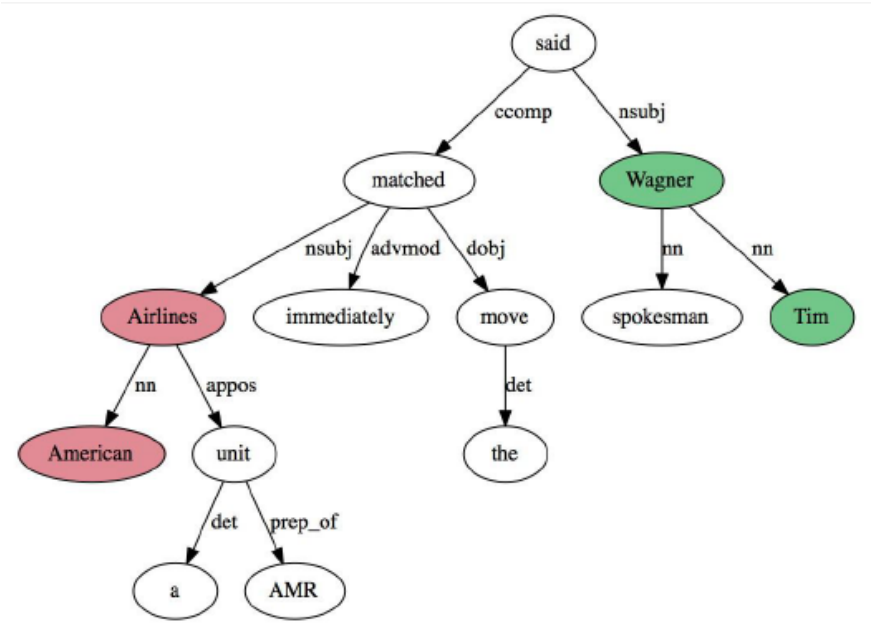
M12-NE = ORG-PERSON

Entity Level of mentions (Name, Nominal, Pronoun)

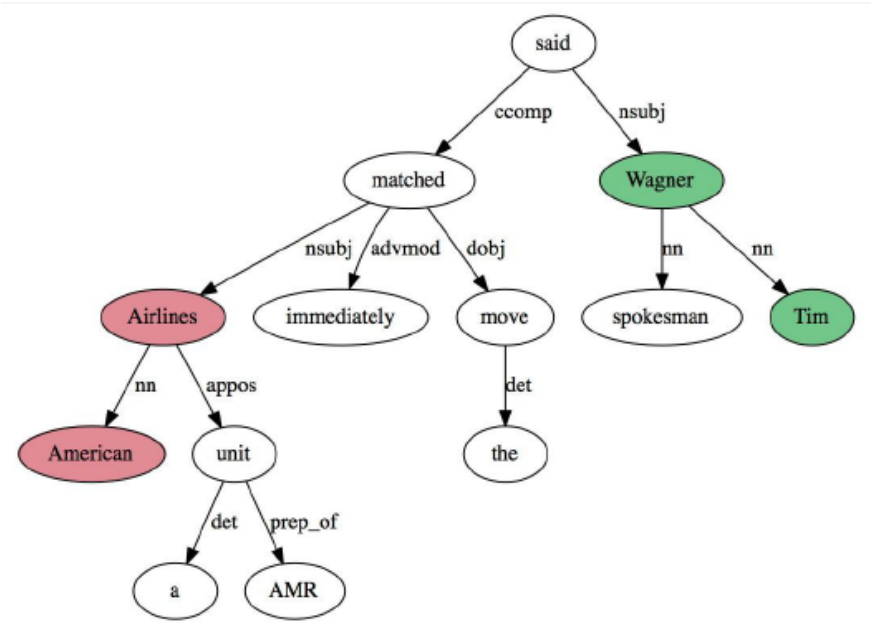
M1:EL = Name, M2:EL = Name

'it' or 'he' would be pronoun, 'the company' would be nominal

Features: dependency syntax features



Features: dependency syntax features



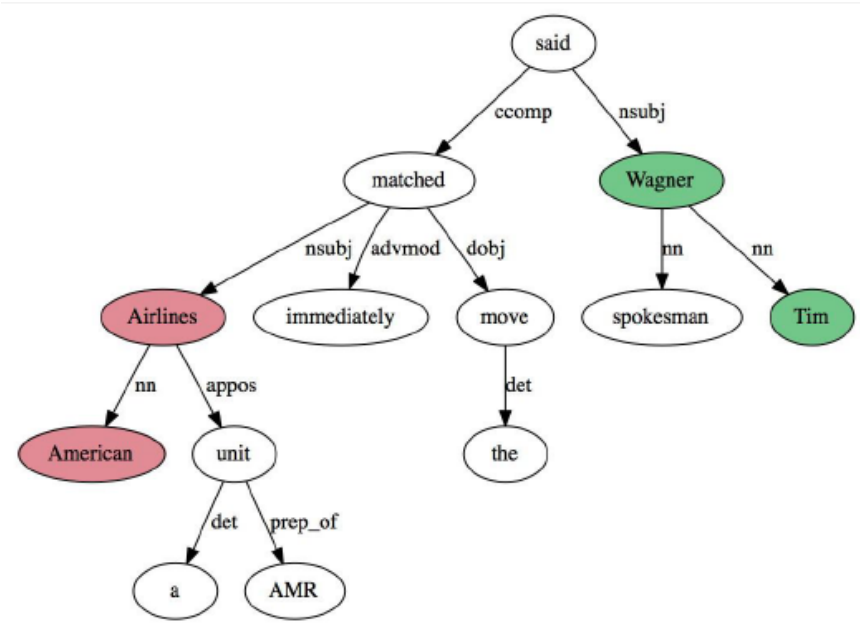
Features of mention dependencies

H1DW1 = matched:Airlines

H2DW2 = said:Wagner

Path = {Airlines, matched, said, Wagner}

Features: dependency syntax features



Features of mention dependencies

H1DW1 = matched:Airlines

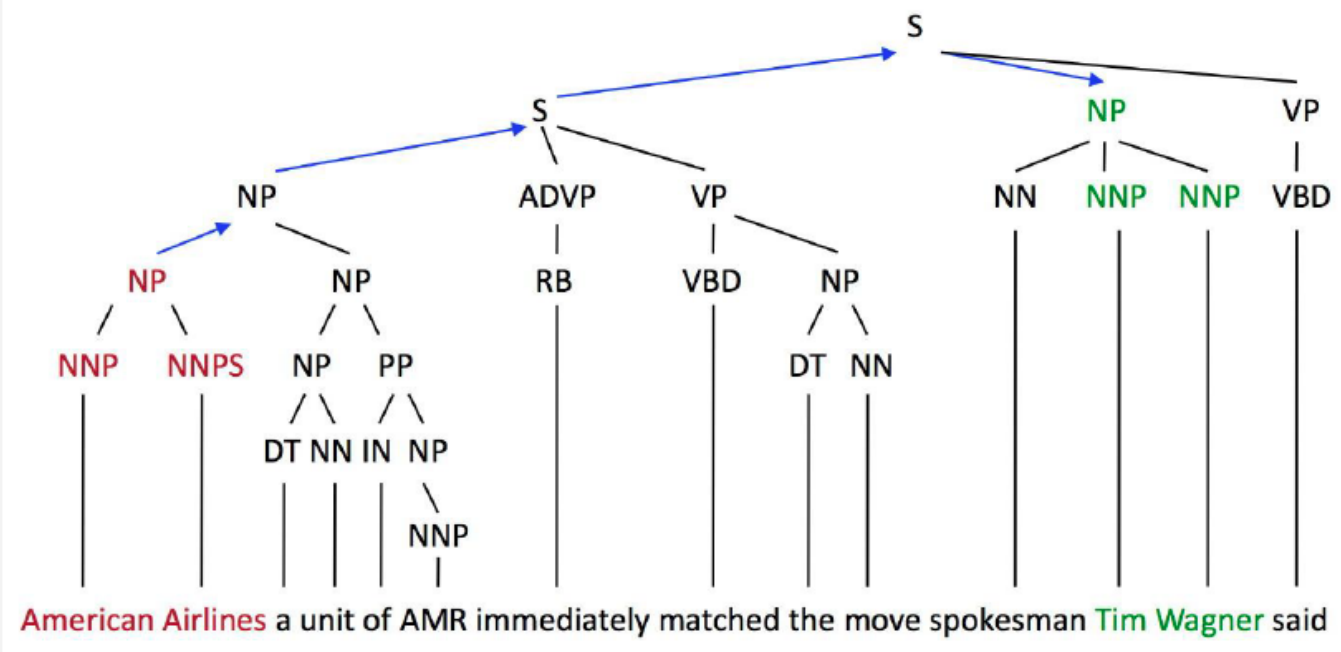
H2DW2 = said:Wagner

Path = {Airlines, matched, said, Wagner}

Base Phrase Chunk Features

[$_{NP}$ American Airlines], [$_{NP}$ a unit] [$_{PP}$ of] [$_{NP}$ AMR], [$_{ADVP}$ immediately] [$_{VP}$ matched] [$_{NP}$ the move], [$_{NP}$ spokesman Tim Wagner] [$_{VP}$ said].

Features: constituency parse features



Features for relation extraction: Gazetteer and trigger word features

Trigger list for family: kinship terms

Features for relation extraction: Gazetteer and trigger word features

Trigger list for family: kinship terms

parent, wife, husband, grandparent etc. [from Wordnet]

Features for relation extraction: Gazetteer and trigger word features

Trigger list for family: kinship terms

parent, wife, husband, grandparent etc. [from Wordnet]

Gazetteer

List of useful geo or geopolitical words

- Country name list
- Other sub-entries

Relation extraction classifiers

Now you can use any classifier

- SVM
- MaxEnt (multiclass logistic regression)
- Naïve Bayes
- etc.

Relation extraction classifiers

Now you can use any classifier

- SVM
- MaxEnt (multiclass logistic regression)
- Naïve Bayes
- etc.

Train it on the training set, tune on the development set, test on the test set

Evaluation of Supervised Relation Extraction

Compute $P/R/F_1$ for each relation

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

Evaluation of Supervised Relation Extraction

Compute $P/R/F_1$ for each relation

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

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

Evaluation of Supervised Relation Extraction

Compute $P/R/F_1$ for each relation

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

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

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

Supervised RE : summary

Supervised approach can achieve high accuracy

- At least, for some relations
- If we have lots of hand-labeled training data

But has significant limitations!

- Labeling large training set (+ named entities) is expensive
- Doesn't generalize to different relations

Supervised RE : summary

Supervised approach can achieve high accuracy

- At least, for some relations
- If we have lots of hand-labeled training data

But has significant limitations!

- Labeling large training set (+ named entities) is expensive
- Doesn't generalize to different relations

Beyond supervised relation extraction

- Distantly supervised relation extraction
- Unsupervised relation extraction