

extraction



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



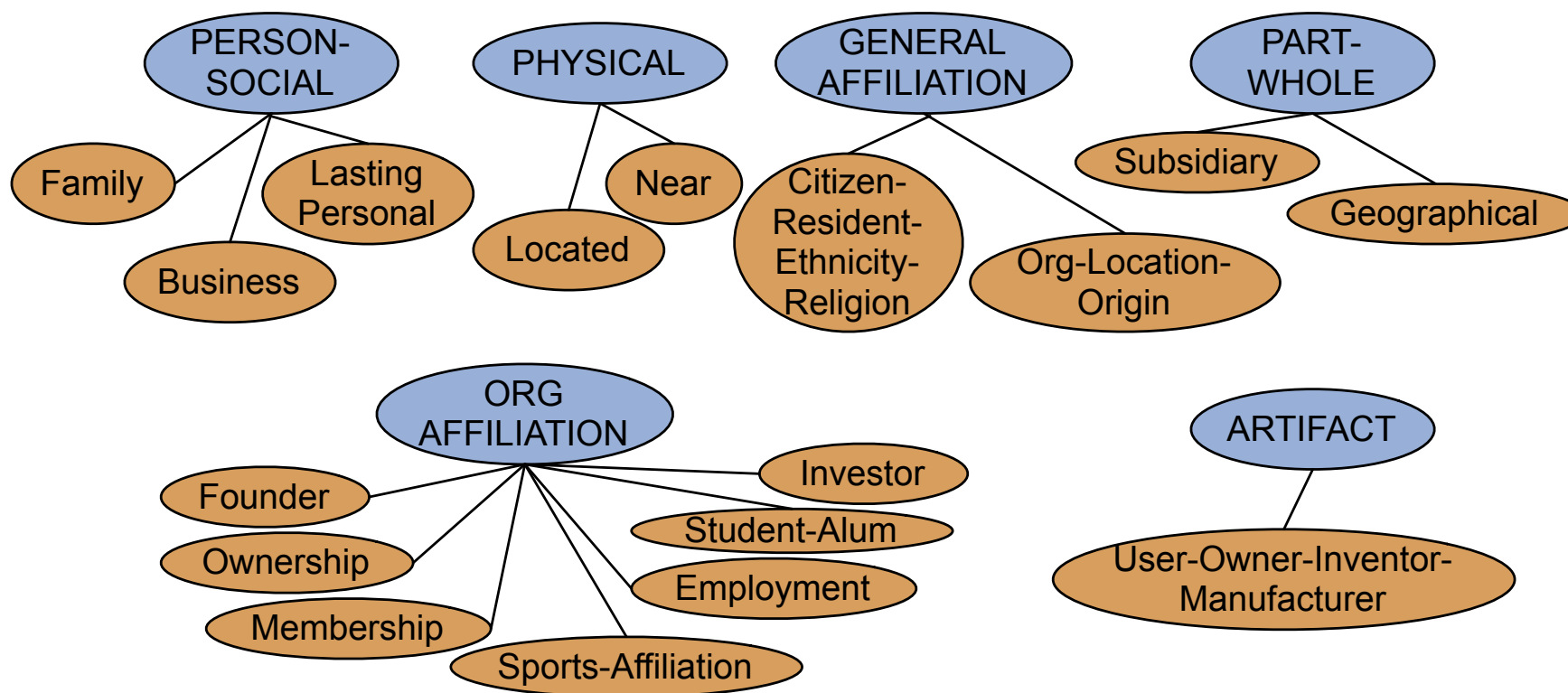
How to do classification in supervised relation extraction

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”

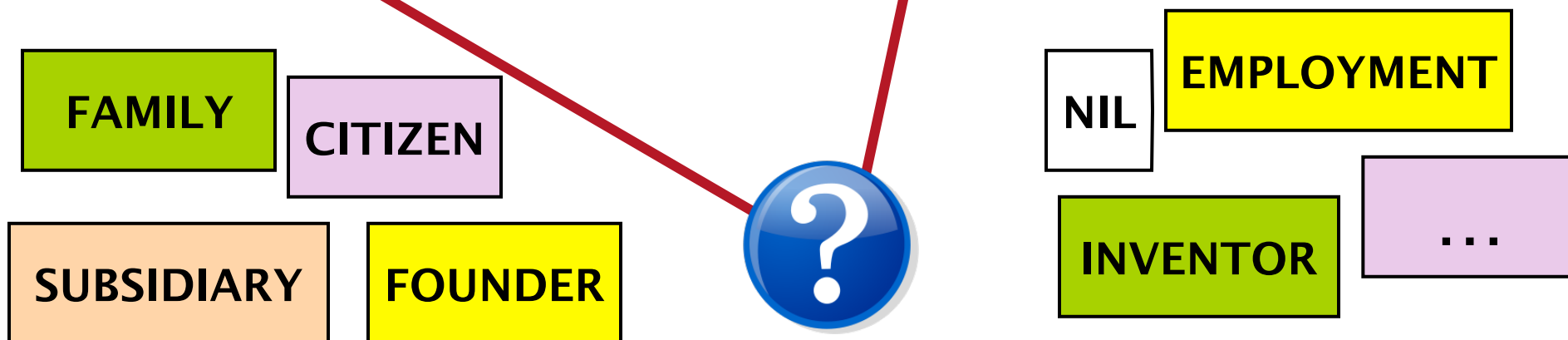




Relation Extraction

Classify the relation between two entities in a sentence

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





Word Features for Relation Extraction

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

Mention 1 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}



Named Entity Type and Mention Level Features for Relation Extraction

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

Mention 1 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**]



Parse Features for Relation Extraction

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

Mention 1 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

Dan Jurafsky



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	<i>airlines</i>
Entity ₂ type	PERS
Entity ₂ 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 ₁	NONE
Word(s) after Entity ₂	<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	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$



Classifiers for supervised methods

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



Evaluation of Supervised Relation Extraction

- Compute $P/R/F_1$ for each relation

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

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

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



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



Supervised relation extraction