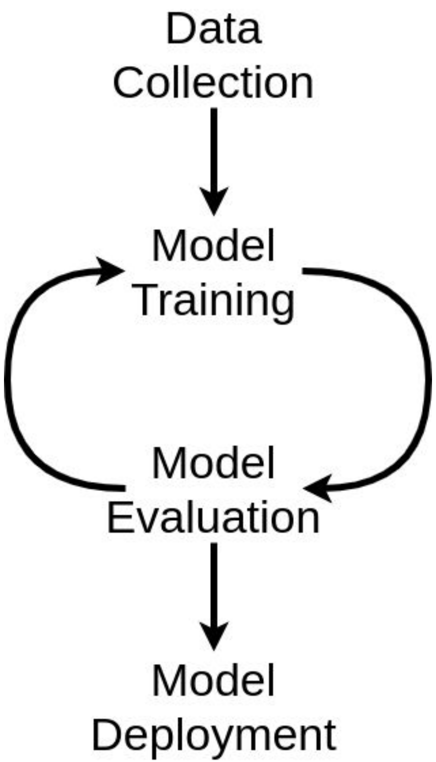
Utilizing extrinsic information in NLP tasks

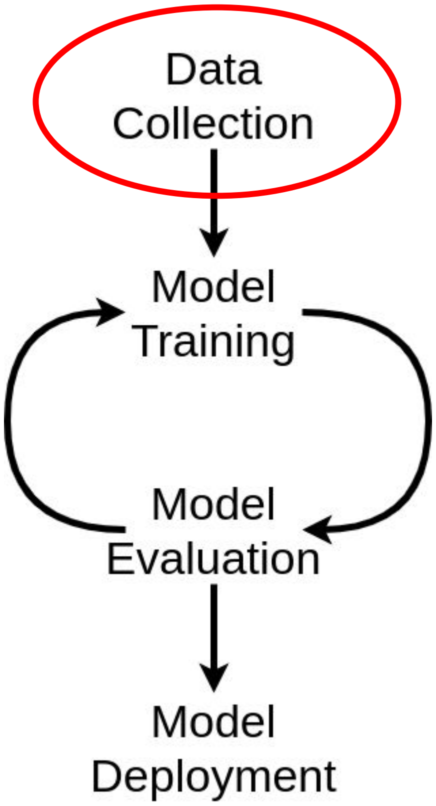
Alon Albalak

Committee: Xifeng Yan (Co-chair), William Yang Wang (Co-chair), Lise Getoor

Machine Learning Pipeline



Machine Learning Pipeline



Data Sources

kaggle

Datasets

GitHub

Data Collection Process

- Label data in-house
  - Time-consuming
  - Domain knowledge
- Crowd-source labels
  - Quality issues
  - Expensive
  - Domain knowledge

How can we build a model with little to no labeled data and limited resources?

Overview

1. Background
2. Dataless Classification
3. Zero-Shot Learning
4. Injecting Domain Knowledge

Overview

1. Background
2. Dataless Classification
3. Zero-Shot Learning
4. Injecting Domain Knowledge

Domains and Tasks

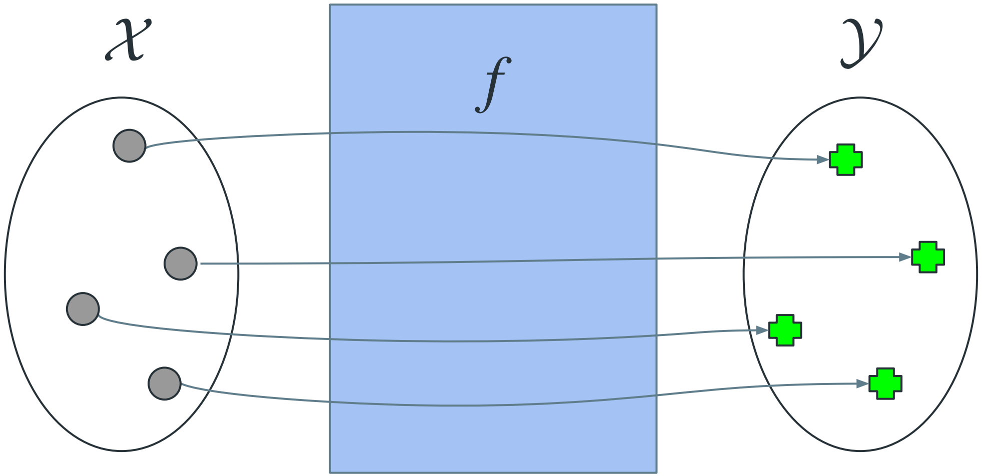
$\mathcal{D}$ - The set of all domains

$\mathcal{T}$ - The set of all tasks

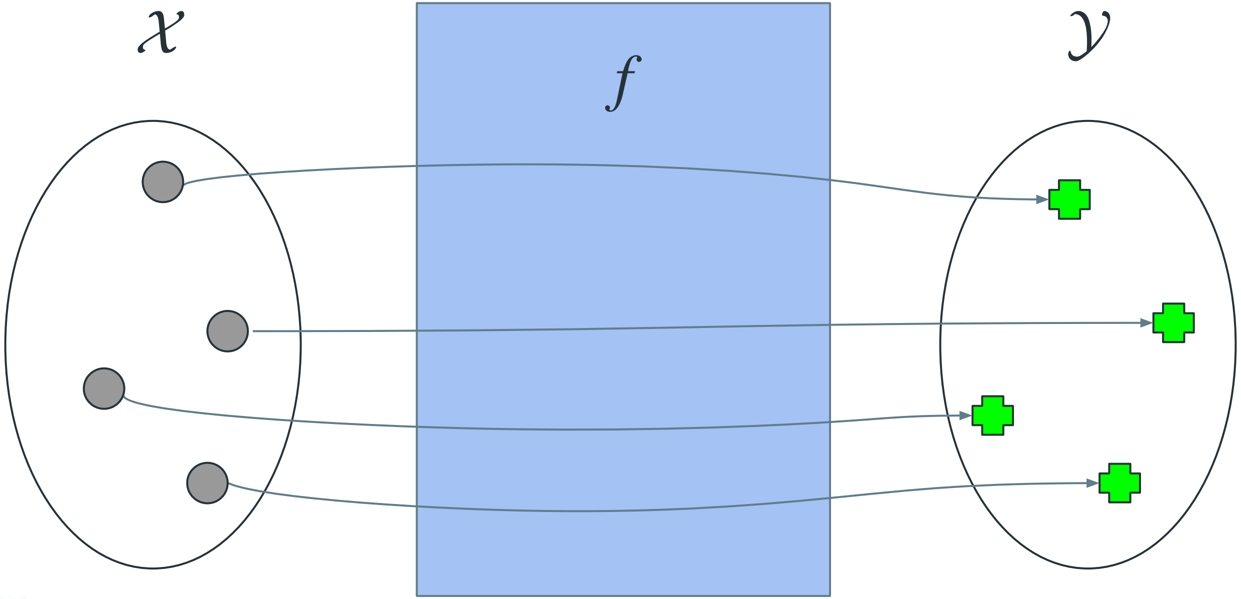
$\mathcal{X}_d$ - Feature space for a given domain, $d \in \mathcal{D}$

$\mathcal{Y}_t$ - Label space for a given task, $t \in \mathcal{T}$

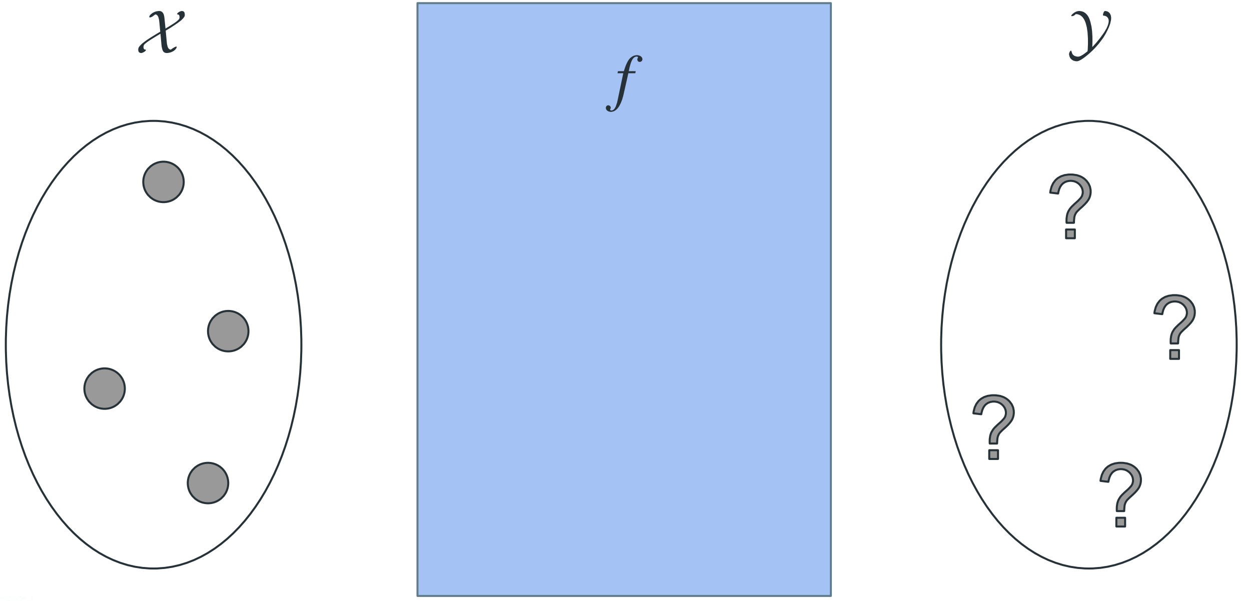
$f : \mathcal{X} \rightarrow \mathcal{Y}$ - A classification function



Supervised Learning



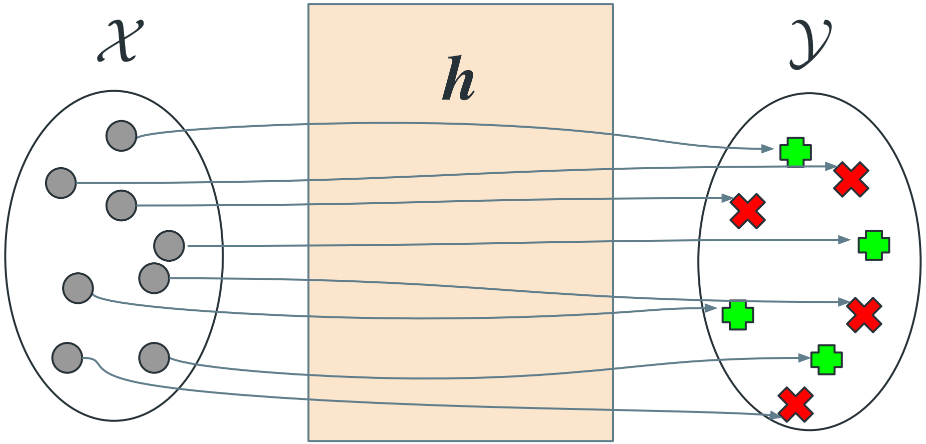
Without labeled data



Weak Supervision

Approximate a mapping function, $\boldsymbol{h}$, using imperfect sources such as:

- Heuristic functions
- Distant supervision
- Semi-supervision



Ratner et al. 2016 | Varma and Re 2018

Relation Extraction

Bill Gates founded Microsoft on April 4th, 1975.

Entity1	Bill Gates
Entity2	Microsoft
Relation	

Mintz et al. 2009 | Riedel et al. 2010

Relation Extraction

Bill Gates founded Microsoft on April 4th, 1975.

Entity1	Bill Gates
Entity2	Microsoft
Relation	FounderOf

Mintz et al. 2009 | Riedel et al. 2010

Distant Supervision in Relation Extraction

WIKIDATA

Entity1	Bill Gates
Entity2	Microsoft
Relation	FounderOf

Mintz et al. 2009 | Riedel et al. 2010

Distant Supervision in Relation Extraction

Bill Gates founded Microsoft on April 4th, 1975.

WIKIDATA

Entity1	Bill Gates
Entity2	Microsoft
Relation	FounderOf

Mintz et al. 2009 | Riedel et al. 2010

Distant Supervision in Relation Extraction

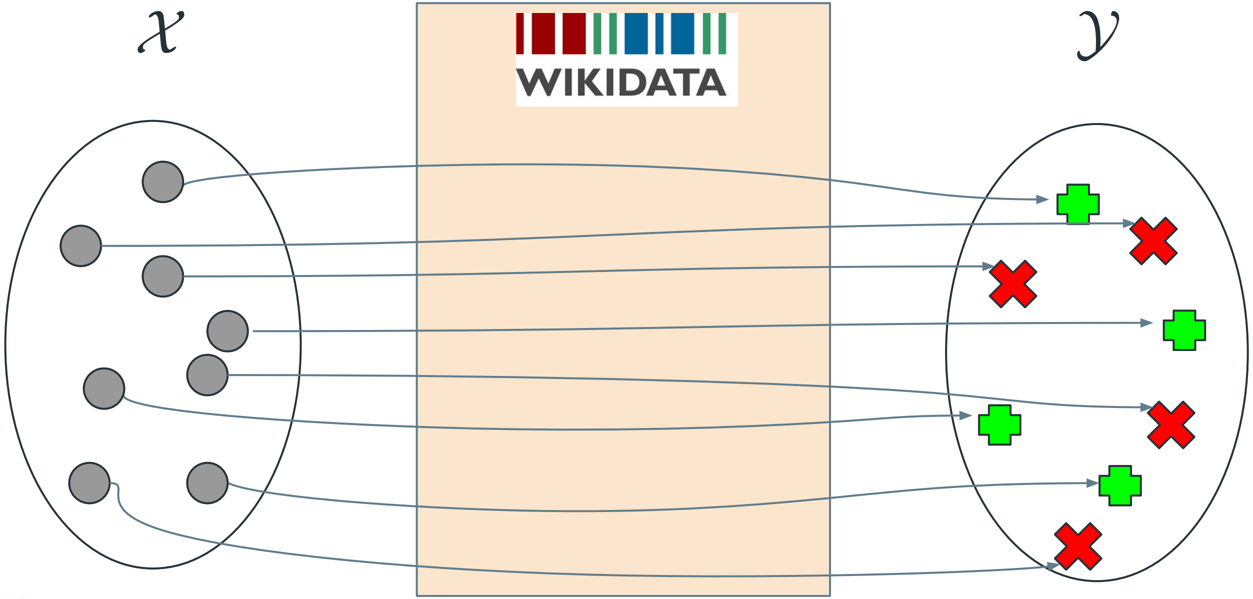
Bill Gates stepped down as CEO of microsoft in 2000.
Bill Gates founded Microsoft on April 4th, 1975.
Bill Gates was the largest shareholder of Microsoft until 2014.

WIKIDATA

Entity1	Bill Gates
Entity2	Microsoft
Relation	FounderOf

Mintz et al. 2009 | Riedel et al. 2010

Distant Supervision in Relation Extraction



Mintz et al. 2009 | Riedel et al. 2010

Transfer Learning

Store knowledge gained from solving one problem, and apply it to a related problem

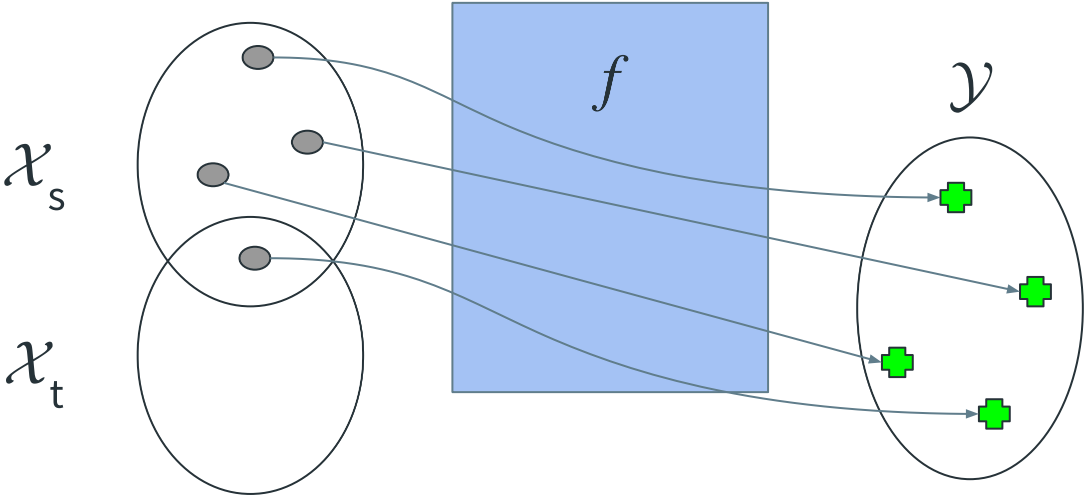
Two types of transfer:

- Domain adaptation
- Task transfer

Pratt et al. 1991 | Ben-David et al. 2009

Domain Adaptation

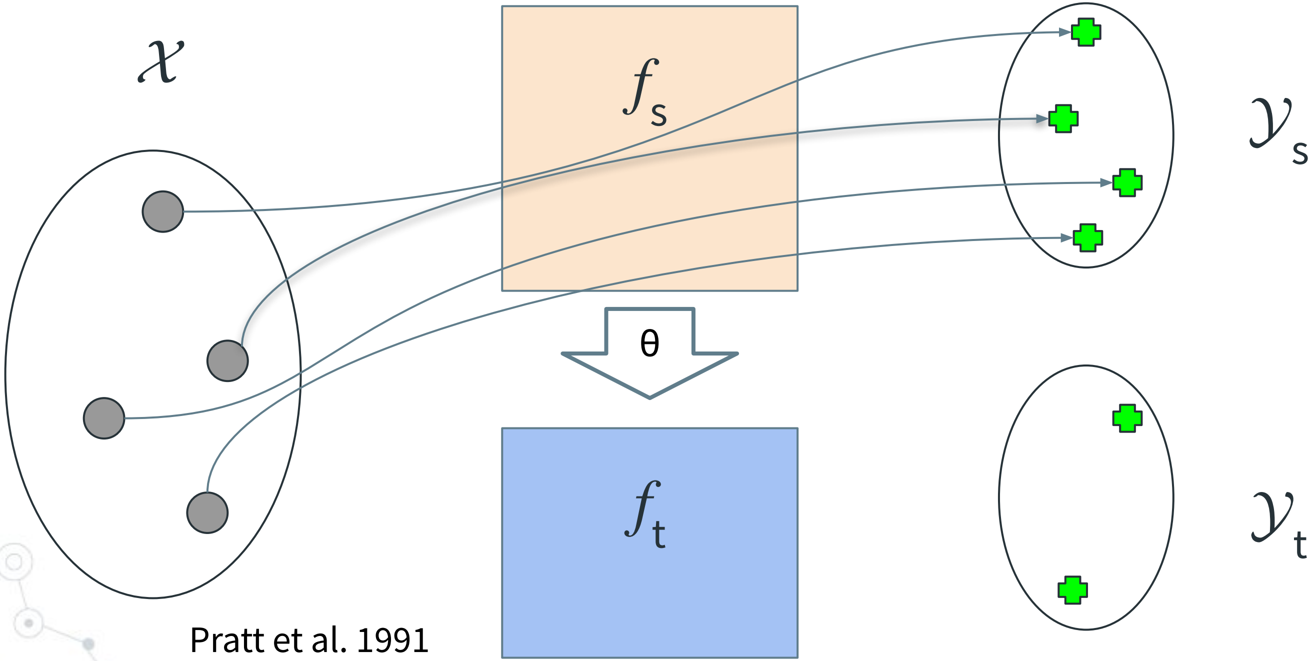
Learn a classifier in a *source* domain, but apply it in a related *target* domain



Ben-David et al. 2009

Task Transfer

Learn a classifier for a *source* task, apply it to a *target* task



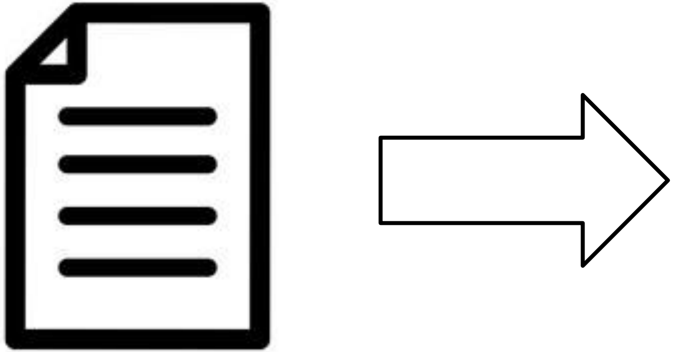
Pratt et al. 1991

Overview

1. Background
2. Dataless Classification
3. Zero-Shot Learning
4. Injecting Domain Knowledge

Importance of Semantic Representation: Dataless Classification

Chang, Ratinov, Roth, and Srikumar. AAAI 2008



Sports
Business
Arts and Humanities
Music

Chang et al. 2008

Dataless Classification: Datasets

- 20 Newsgroups Dataset:
  - Discussion forums
  - Categories such as:
    - Talk.politics.mideast
    - Sci.electronics
    - Rec.sport.hockey
- Yahoo! Answers Dataset:
  - Question/Answer pairs
  - Categories and subcategories such as:
    - Arts and Humanities - Theater Acting
    - Sports - Rugby League

Chang et al. 2008

Text Classification

$\mathcal{D}$ - Discussion Forums, Question/Answer pairs

$\mathcal{X}$ - Document text

$\mathcal{Y}$ - Newsgroup names, Question/Answer categories

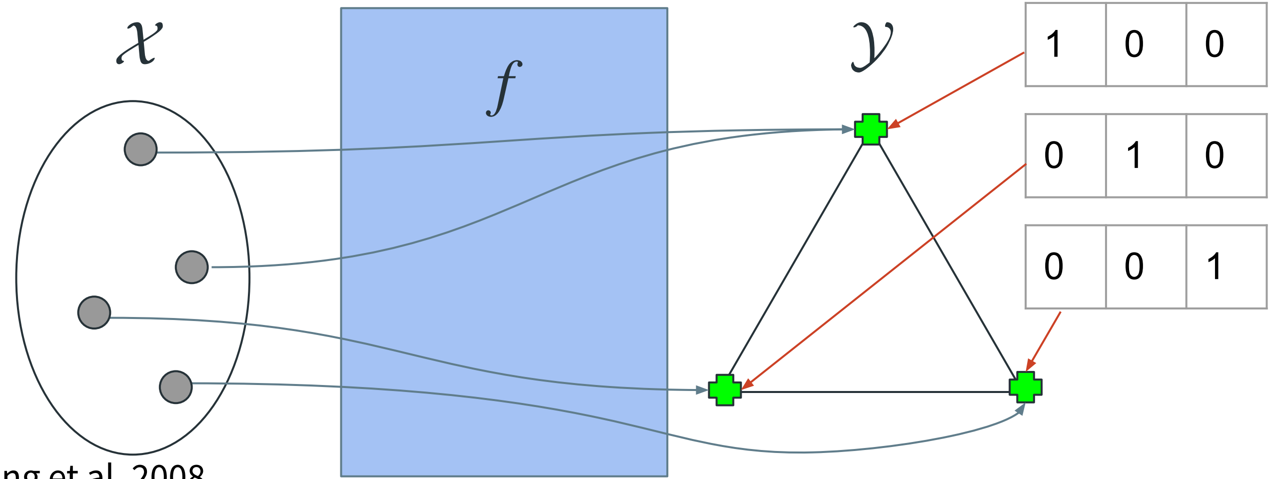
Chang et al. 2008

Text Classification

$\mathcal{D}$ - Discussion Forums, Question/Answer pairs

$\mathcal{X}$ - Document text

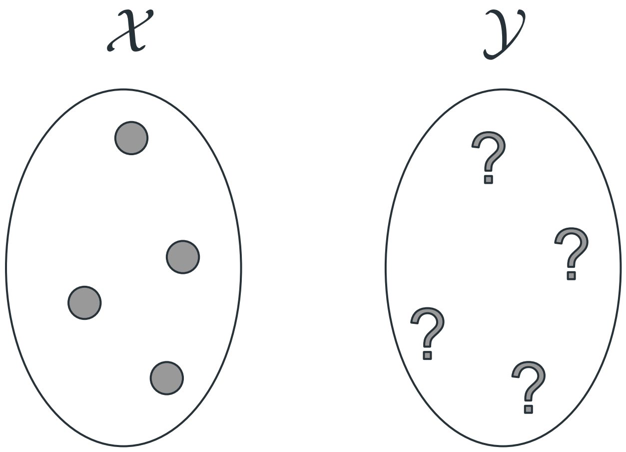
$\mathcal{Y}$ - Newsgroup names, Question/Answer categories



Chang et al. 2008

On-the-Fly Text Classification

- Given the full text corpus
- No labels associated with each text
- Label names only given at test time

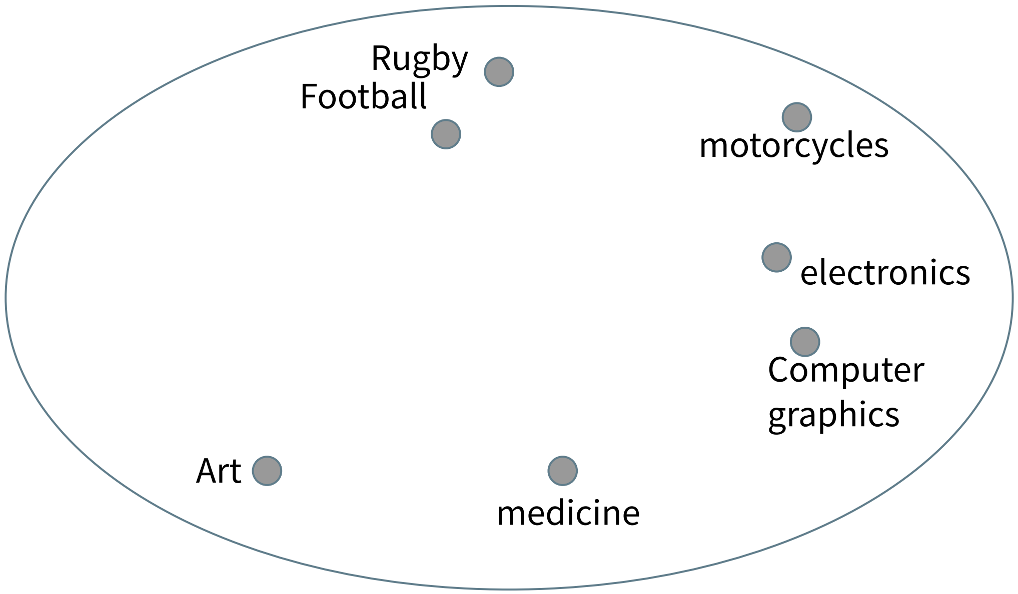


Chang et al. 2008

People can categorize documents into named categories without any explicit training because we know the meaning of category names

Chang et al. 2008

Dataless Classification: Semantic Representation



Chang et al. 2008

Dataless Classification: Semantic Representation

2 methods of creating a semantic space:

- Bag-of-words
- Explicit Semantic Analysis

Bag-of-words model:

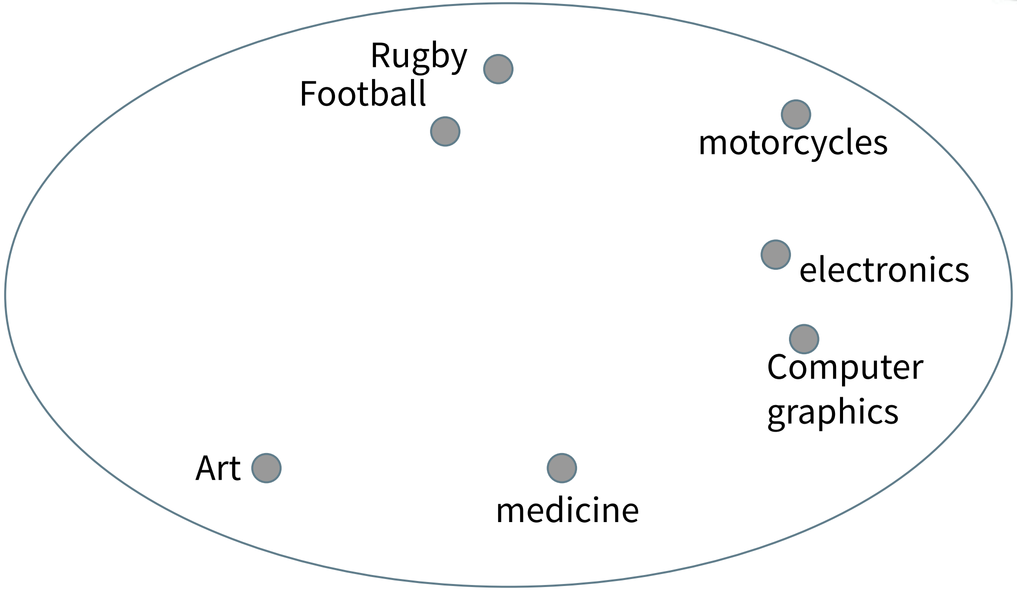
"I enjoyed the movie…" =

best	enjoyed	I	movie	the
0	1	1	1	1

Chang et al. 2008

Dataless Classification: On-the-fly Classification

Category names are not known in advance

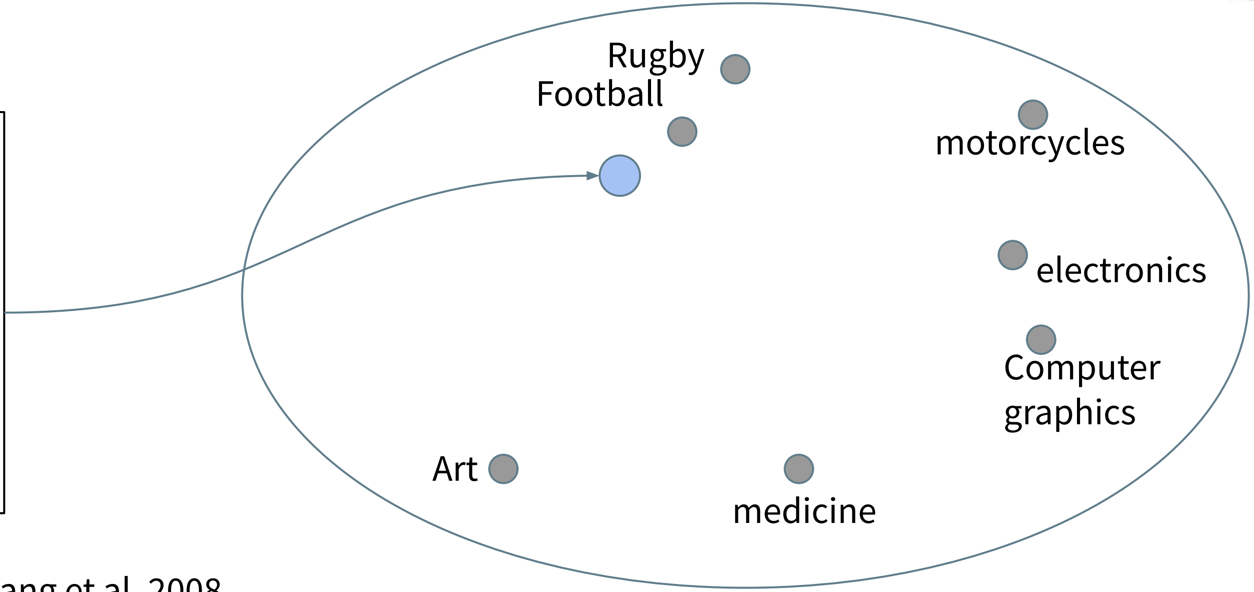


Chang et al. 2008

Dataless Classification: On-the-fly Classification

Category names are not known in advance

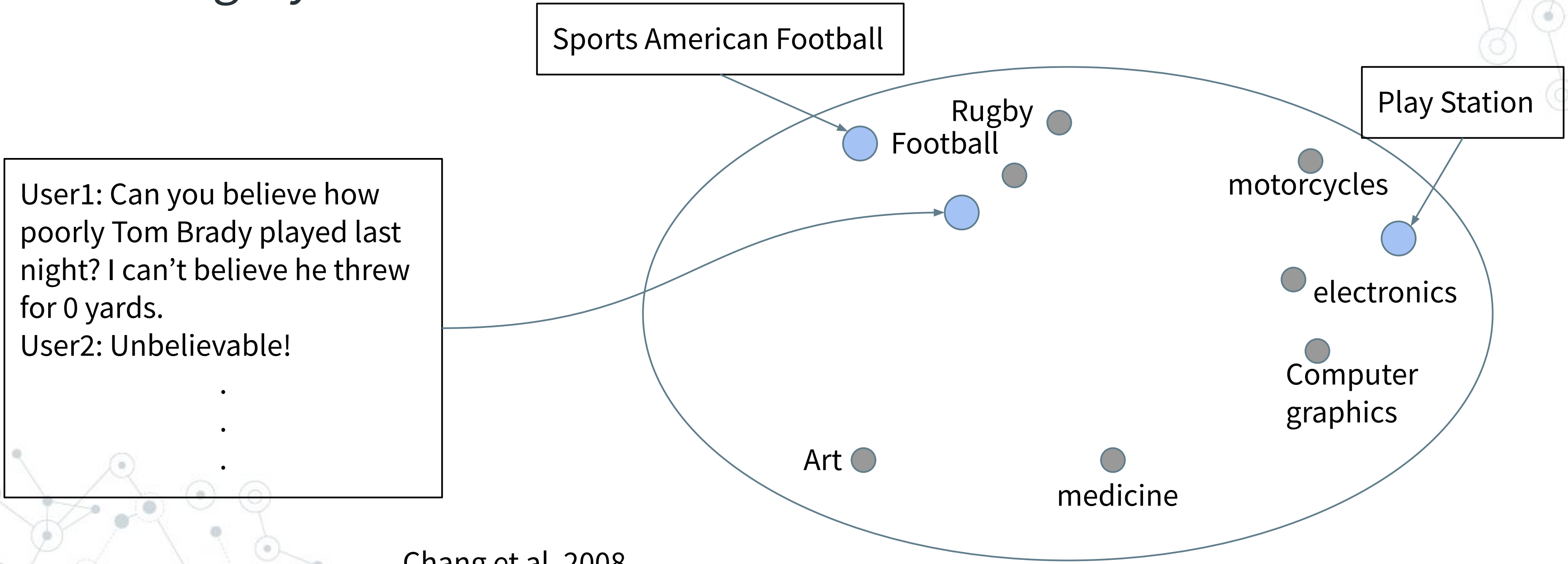
User1: Can you believe how poorly Tom Brady played last night? I can't believe he threw for 0 yards.
User2: Unbelievable!
.
.
.



Chang et al. 2008

Dataless Classification: On-the-fly Classification

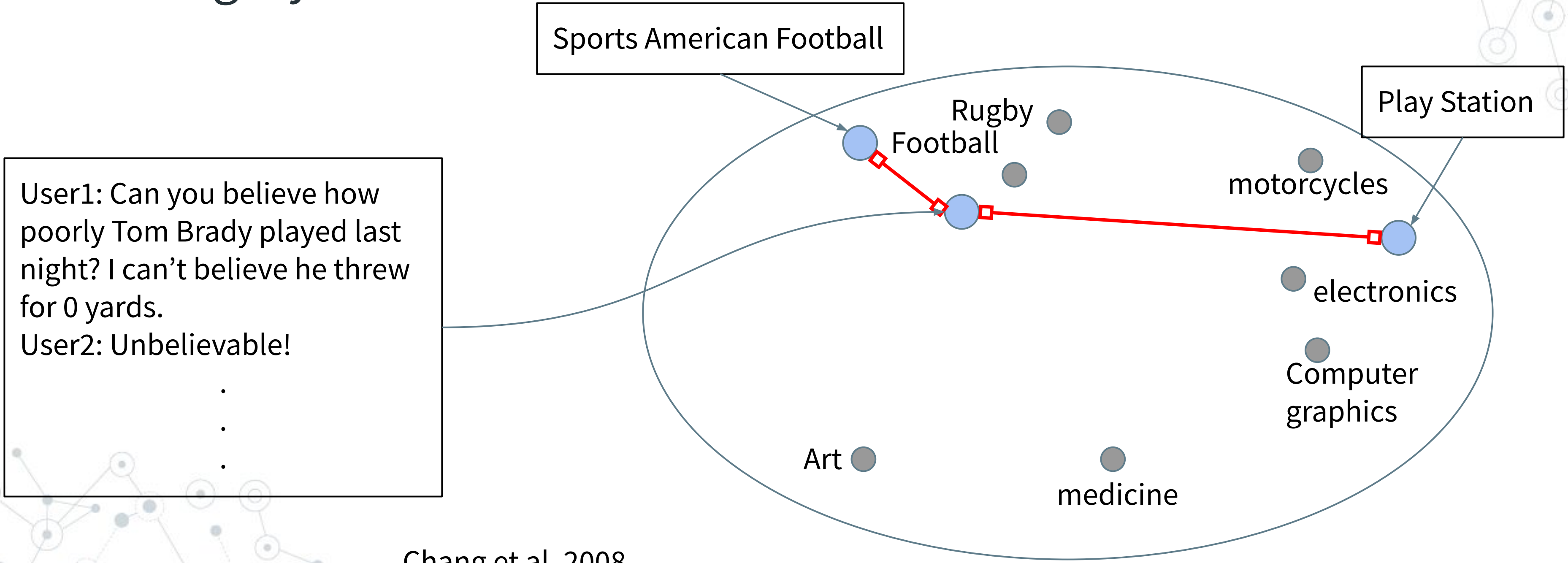
Category names are not known in advance



Chang et al. 2008

Dataless Classification: On-the-fly Classification

Category names are not known in advance



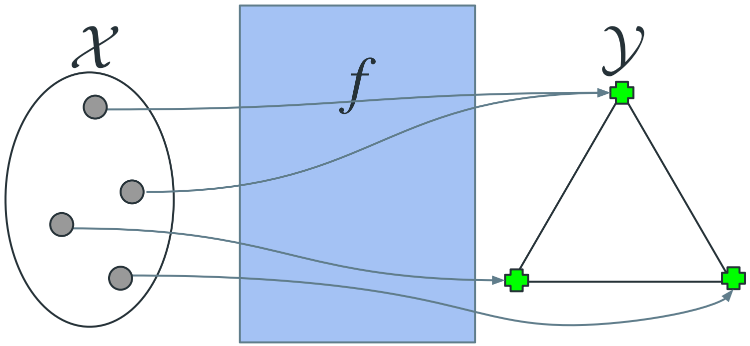
Chang et al. 2008

Dataless Classification: Results

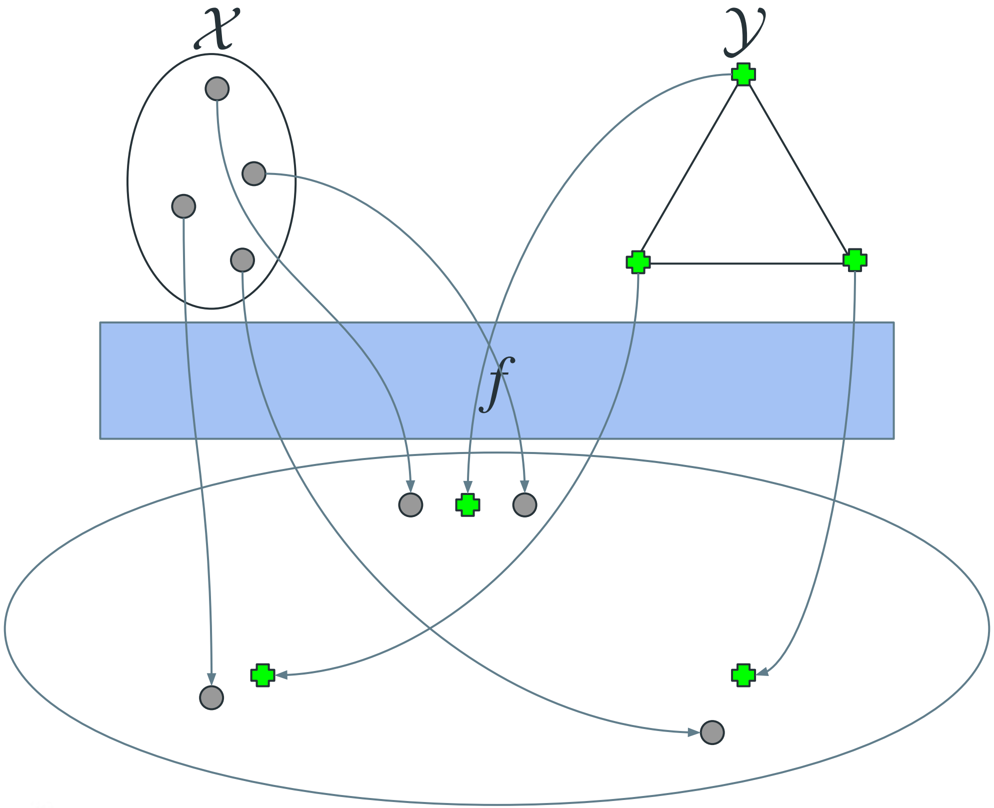
Dataset / Accuracy	Supervised Baseline (10)	Supervised Baseline (100)	Nearest Neighbors Bag-of-Words	Nearest Neighbors ESA
Newsgroups	71.71	92.41	65.73	85.29
Yahoo! Answers	84.34	94.37	66.79	88.62

Chang et al. 2008

Dataless Classification: Takeaway



Dataless Classification: Takeaway



Overview

1. Background
2. Dataless Classification
3. Zero-Shot Learning
4. Injecting Domain Knowledge

Zero-Shot Relation Extraction via Reading Comprehension

Levy, Seo, Choi, and Zettlemoyer. CoNLL 2017

Relation Extraction:

- Given a text and entity pair, determine the relation

Relation Extraction - Slot Filling:

- Given a text, entity, and relation, determine the target entity from the text

Zero-shot Relation Extraction - Slot filling:

- Test-time relations are unobserved during training

Levy et al. 2017

Zero-Shot Relation Extraction

$\mathcal{D}$ - WikiReading

$\mathcal{T}$ - Relation Extraction

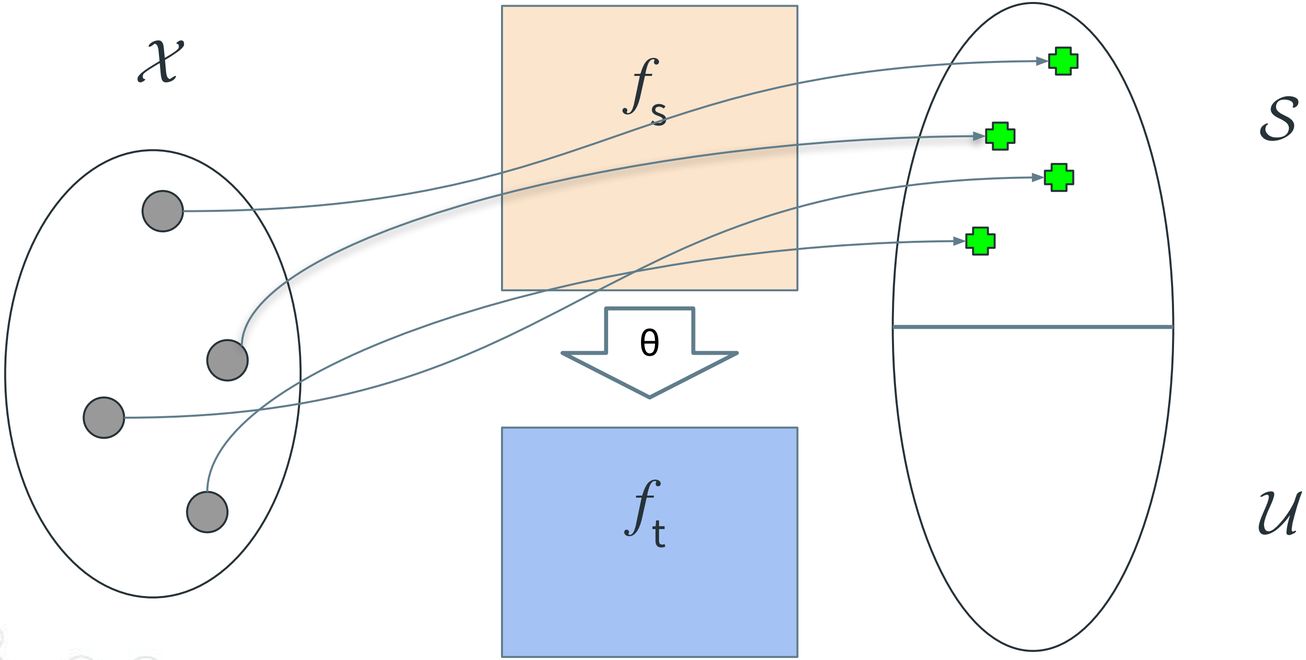
$\mathcal{X}$ - Document text

$\mathcal{Y}$ - $\mathcal{S} \cup \mathcal{U}, \mathcal{S} \cap \mathcal{U} = \varnothing$

- $\mathcal{S}$ - Seen relations (eg. founderOf, educatedAt, etc.)
- $\mathcal{U}$ - Unseen relations (eg. occupation, spouse, etc.)

Levy et al. 2017

Zero-Shot Relation Extraction



Levy et al. 2017

Zero-Shot Relation Extraction: Task Reformulation

$\mathcal{D}$ - WikiReading

$\mathcal{T}$ - Reading Comprehension (Span extraction)

$\mathcal{X}$ - Document text

$\mathcal{Y}$ - Start/end tokens

Levy et al. 2017

Zero-Shot Relation Extraction: Task Reformulation

$\mathcal{D}$ - WikiReading

$\mathcal{T}$ - Reading Comprehension (Span extraction)

$\mathcal{X}$ - Document text

$\mathcal{Y}$ - Start/end tokens

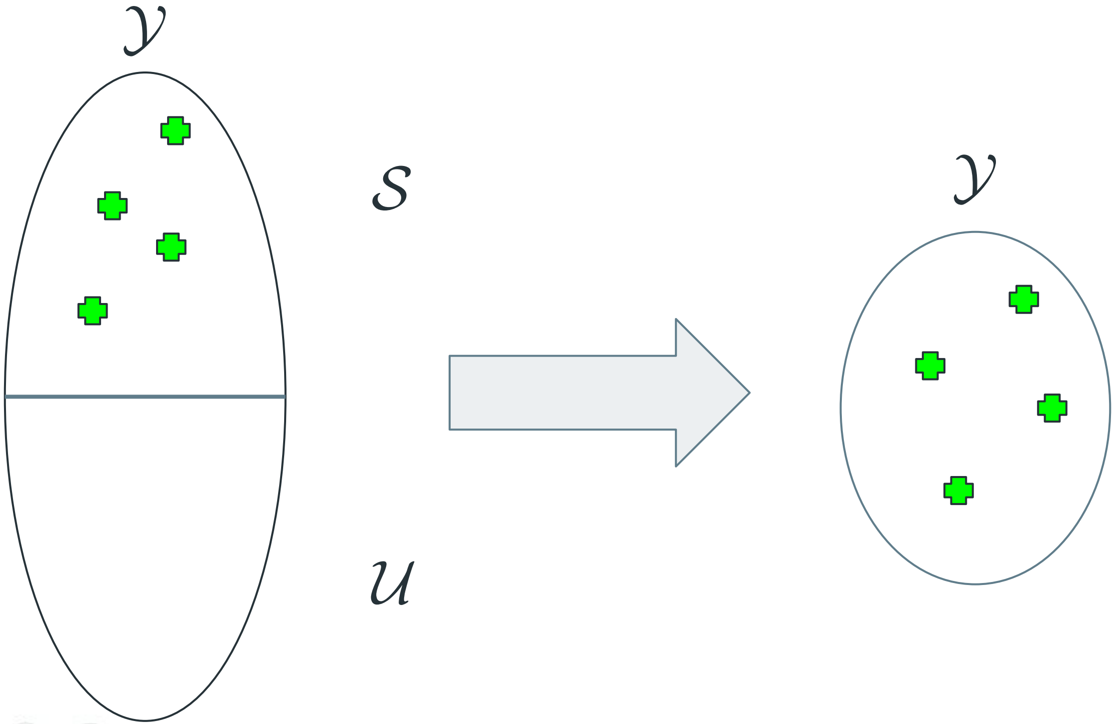
Question: Who founded Microsoft?

Context: Bill Gates founded Microsoft on April 4th, 1975.

Answer: Bill Gates founded Microsoft on April 4th, 1975.

Levy et al. 2017

Zero-Shot Relation Extraction as Reading Comprehension



Levy et al. 2017

Zero-Shot Relation Extraction: Querification

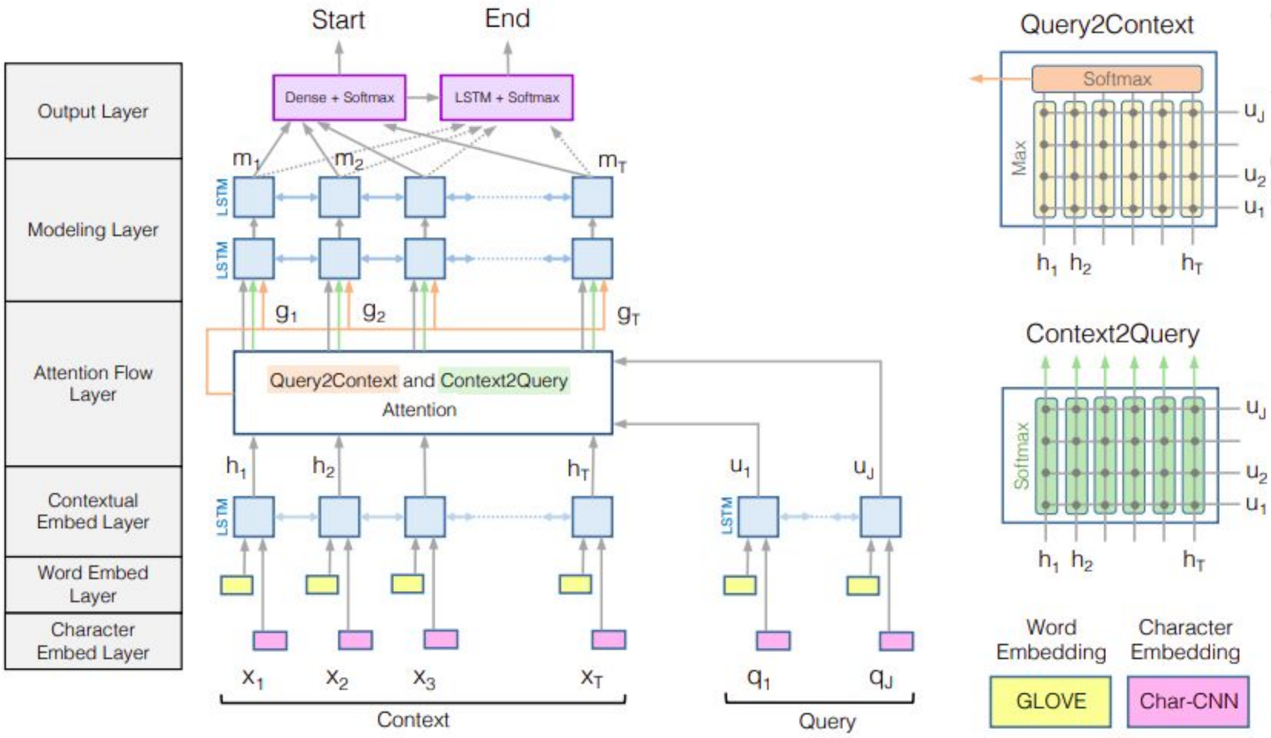
Each relation must be turned into a question template:

- EducatedAt → Where did X go to school?
- Spouse → Who is the spouse of X?

Levy et al. 2017

Zero-Shot Relation Extraction: Reading Comprehension

BiDAF Model



Seo et al. 2016

Zero-Shot Relation Extraction: Results

Method	Precision	Recall	F1
Random Named Entity	9.25	18.06	12.23
RNN labeler	13.28	5.69	7.97
Single Question	37.18	31.24	33.9
Question Ensemble	45.85	37.44	41.11

Levy et al. 2017

Zero-Shot Relation Extraction: Takeaways

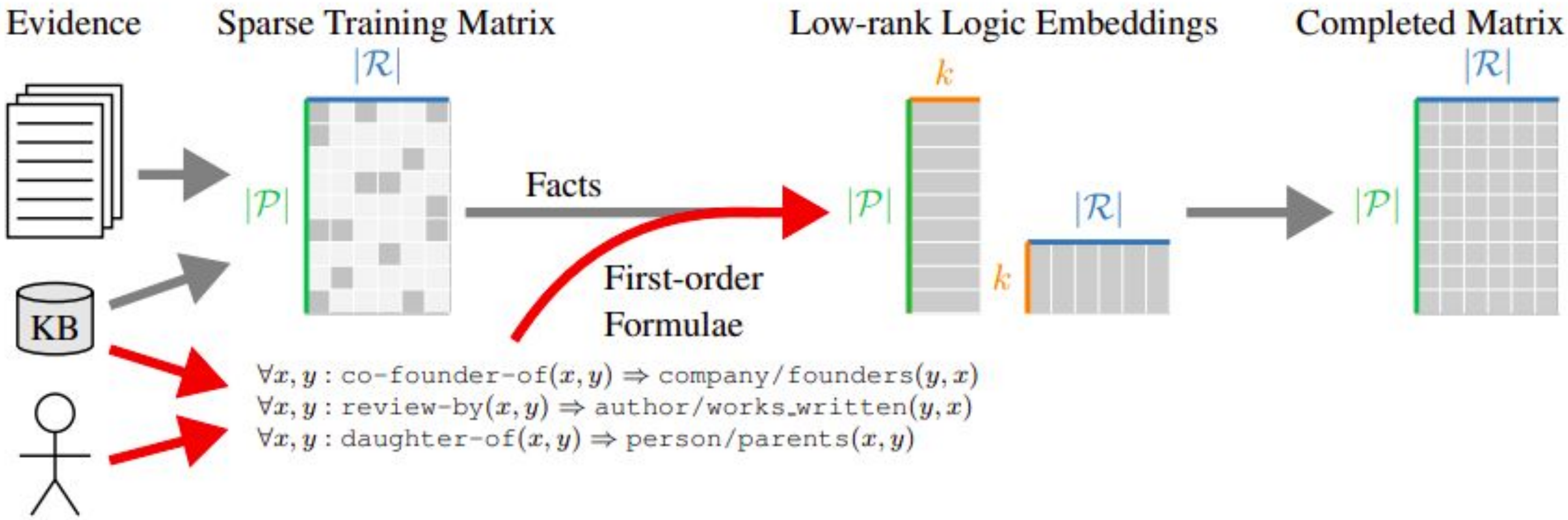
- External information in the form of natural language questions
- Relies heavily on good semantic representation

Overview

1. Background
2. Dataless Classification
3. Zero-Shot Learning
4. Injecting Domain Knowledge

Injecting Logical Background Knowledge into Embeddings for Relation Extraction

Rocktaschel, Singh, and Riedel. NAACL-HTL 2015



Rocktaschel et al. 2015

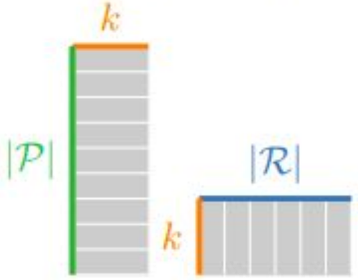
Matrix Factorization in Relation Extraction

- $\mathcal{E}$ - The set of entities
- $\mathcal{R}$ - The set of relations
- $\mathcal{P} \subseteq \mathcal{E} \times \mathcal{E}$ - The set of entity-pairs
- Knowledge base matrix:
  - $\mathcal{P} \times \mathcal{R}$ matrix
  - Rows represent entity-pairs
  - Columns represent relations
- Goal:
  - Find a low rank factorization of the knowledge base matrix
  - Embedding matrix of entity-pairs - $|\mathcal{P}| \times k$
  - Embedding matrix of relations - $k \times |\mathcal{R}|$

	FounderOf
BillGates Microsoft	1
ElonMusk Microsoft	0

Rocktaschel et al. 2015

Matrix Factorization in Relation Extraction



- $\mathbf{v}_r$ - embedding for relation r
- $\mathbf{v}_{(i,j)}$ - embedding for entity pair (e_i,e_j)
- $\pi_r^{(i,j)} = \sigma(\mathbf{v}_r \cdot \mathbf{v}_{(i,j)})$ - probability of relation r holding for entity pair (e_i,e_j)
- Conditional probability of a possible world **w** with the embeddings **V** is: $p(\mathbf{w}|\mathbf{V}) = \prod_{r(i,j)\in w} \pi_r^{(i,j)} \prod_{r(i,j)\notin w} (1 - \pi_r^{(i,j)})$
- Loss: $-\log(p(\mathbf{w}|\mathbf{V}))$

Rocktaschel et al. 2015

Injecting Logic into Factorization: Pre-factorization Inference

Add inferred facts as additional training data

Eg: founderOf(x,y) ⇒ workedAt(x,y)

Rocktaschel et al. 2015

Injecting Logic into Factorization: Pre-factorization Inference

Add inferred facts as additional training data

Eg: founderOf(x,y) ⇒ workedAt(x,y)

	founderOf	workedAt
BillGates Microsoft	1	0
Elon Musk Microsoft	0	0

	founderOf	workedAt
BillGates Microsoft	1	1
Elon Musk Microsoft	0	0

Rocktaschel et al. 2015

Injecting Logic into Factorization: Joint Optimization

- $\mathcal{F}$ - logical formula
- $[\mathcal{F}]$ - probability, p($\mathbf{w}|\mathbf{V}$), that $\mathcal{F}$ is true
- $[\mathcal{A} \vee \mathcal{B}] = [\mathcal{A}] + [\mathcal{B}] - [\mathcal{A}][\mathcal{B}]$
- $[\mathcal{A} \Rightarrow \mathcal{B}] = [\mathcal{A}]([\mathcal{B}] - 1) + 1$
- $[\forall x,y: \mathcal{F}] = \prod_{x,y} [\mathcal{F}]$

- $$\min_{\mathbf{v}} \sum_{\mathcal{F}} -\log(\mathcal{F})$$

Rocktaschel et al. 2015

Injecting Logic into Factorization: Evaluation

Using distant supervision from Freebase

Fill matrix with textual patterns from NY-Times corpus

Train on:

- 3960 textual patterns for 139 relations
- 41913 entity-pairs
- 111488 facts

Test on:

- 12 relations that have no textual pattern
- 7293 facts

Rocktaschel et al. 2015

Injecting Logic into Factorization: Methods

- **MF** - Matrix Factorization
- **INF** - Logical Inference
- **Post** - Post-factorization Inference
- **Pre** - Pre-factorization Inference
- **Joint** - Joint optimization

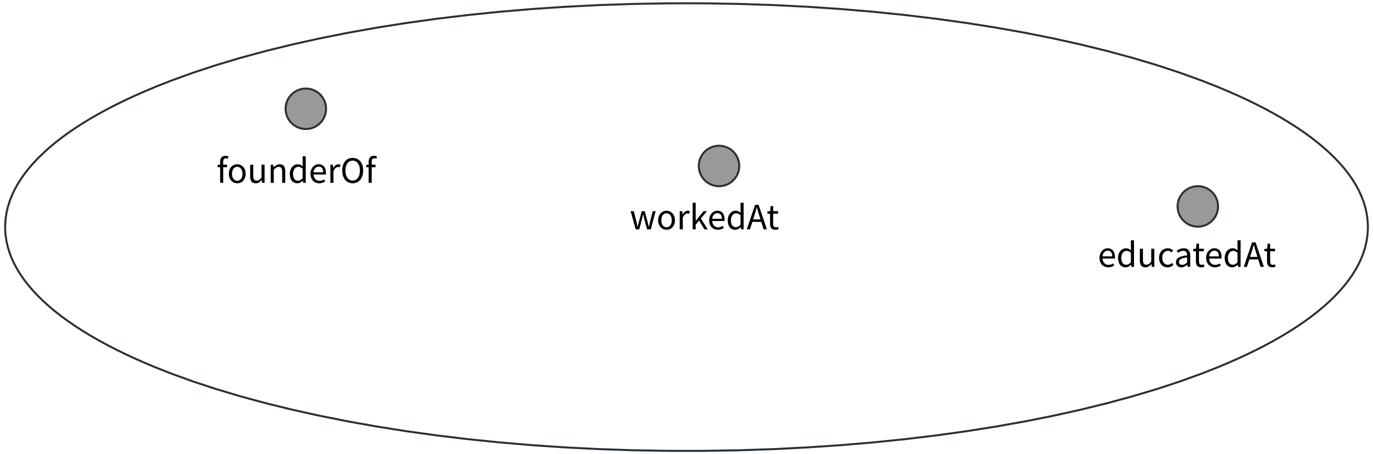
Injecting Logic into Factorization: Results

Zero-shot Relation Extraction	**MF**	**Inf**	**Post**	**Pre**	**Joint**
MAP	0.01	0.23	0.34	0.43	0.52
wMAP	0.03	0.1	0.21	0.33	0.38

Rocktaschel et al. 2015

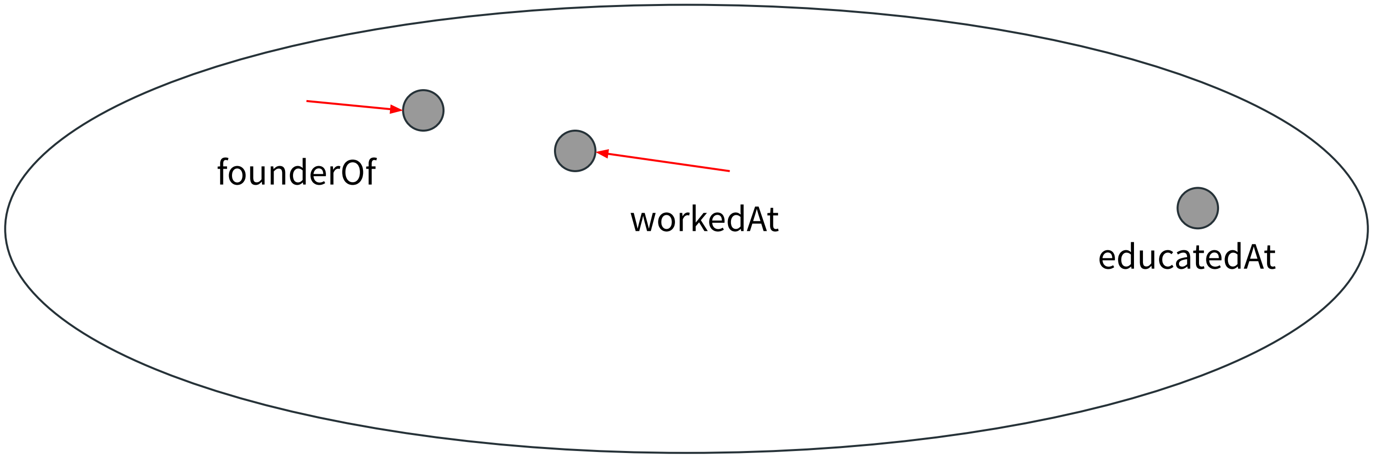
Injecting Logic into Factorization: Takeaway

founderOf(x,y) ⇒ workedAt(x,y)



Injecting Logic into Factorization: Takeaway

$$\text{founderOf}(x,y) \Rightarrow \text{workedAt}(x,y)$$



Thank you!
Questions?