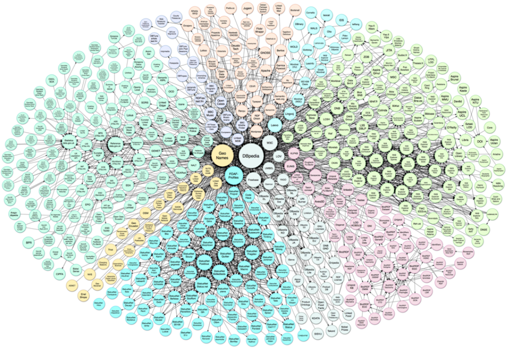
Semantic labeling
A domain-independent approach

Minh Pham, Suresh Alse, Craig Knoblock, Pedro Szekely

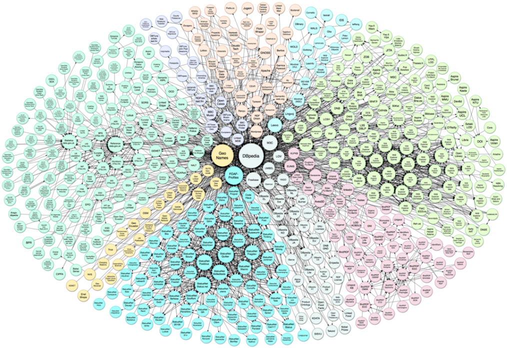
Information Science Institute
University of Southern California

October 21, 2016

How can we integrate data ?

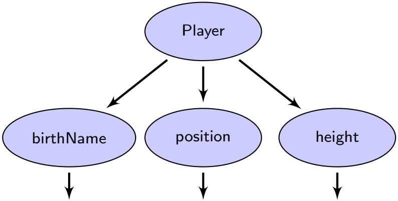


How can we integrate data ?



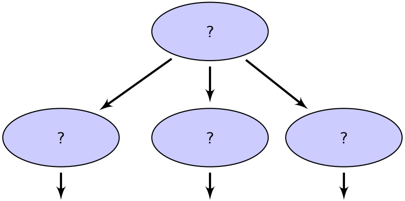
SEMANTIC LABELING

What is Semantic Labeling ?



full name	position	height
Hazard, Eden	Midfielder	172
Cahill, Gary	Defender	191
Rooney, Wayne	Forward	176

Labeled source



fn	ht	ps
Alan PULIDO	176 cm	FW
Robin VAN PERSIE	186 cm	MF
Miiko ALBORNOZ	180 cm	DF

Unlabeled source

Outline

1 Previous approach: domain-dependent

2 Our approach: domain-independent

3 Similarity features

4 Evaluation

5 Conclusion and Future Work

Outline

1 Previous approach: domain-dependent

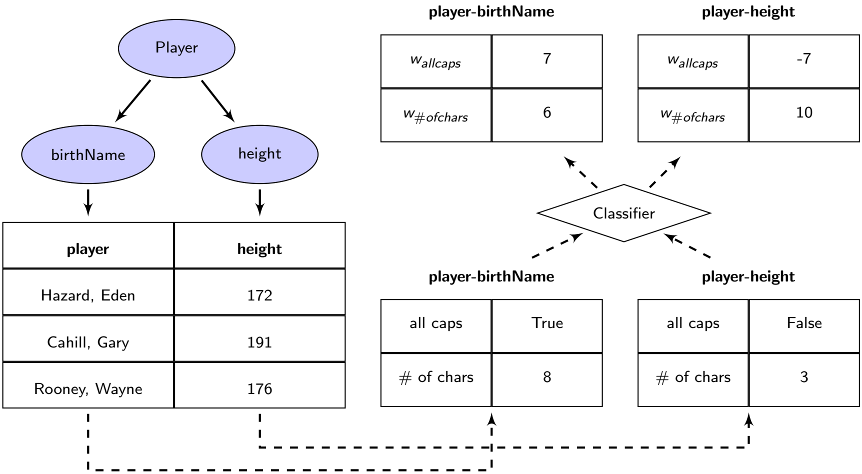
2 Our approach: domain-independent

3 Similarity features

4 Evaluation

5 Conclusion and Future Work

Domain-dependent approach: Training



player	height
Hazard, Eden	172
Cahill, Gary	191
Rooney, Wayne	176

player-birthName

$w_{allcaps}$	7
$w_{\#ofchars}$	6

player-height

$w_{allcaps}$	-7
$w_{\#ofchars}$	10

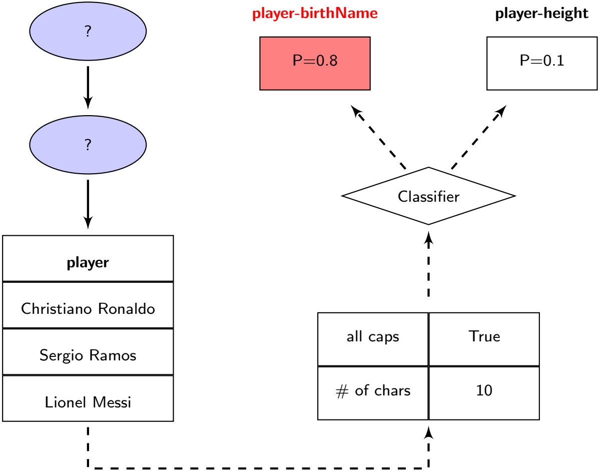
player-birthName

all caps	True
# of chars	8

player-height

all caps	False
# of chars	3

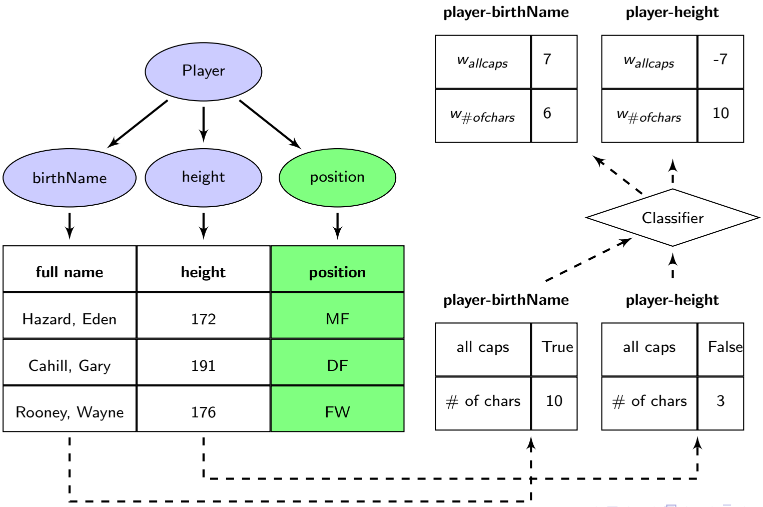
Domain-dependent approach: Predicting



player
Christiano Ronaldo
Sergio Ramos
Lionel Messi

all caps	True
# of chars	10

Domain-dependent approach: Adding new attribute



full name	height	position
Hazard, Eden	172	MF
Cahill, Gary	191	DF
Rooney, Wayne	176	FW

player-birthName

$w_{allcaps}$	7
$w_{\#ofchars}$	6

player-height

$w_{allcaps}$	-7
$w_{\#ofchars}$	10

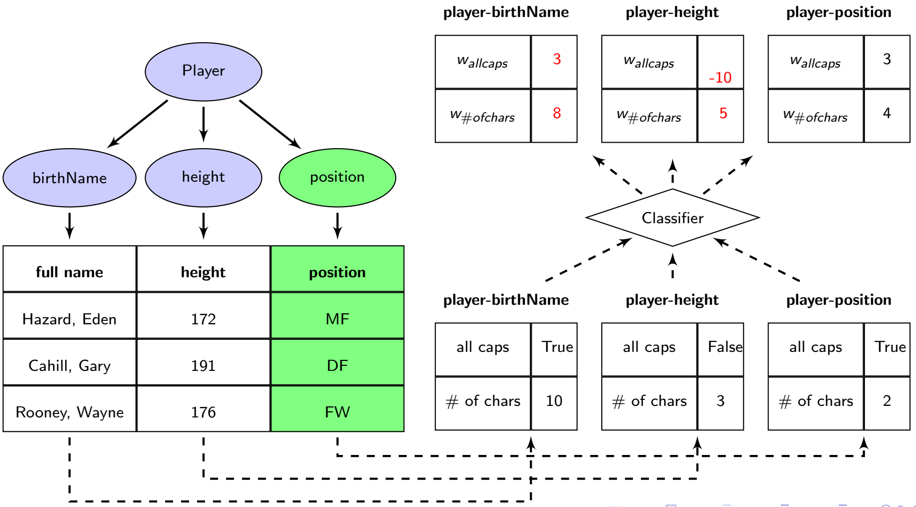
player-birthName

all caps	True
# of chars	10

player-height

all caps	False
# of chars	3

Domain-dependent approach: Adding new attribute



full name	height	position
Hazard, Eden	172	MF
Cahill, Gary	191	DF
Rooney, Wayne	176	FW

player-birthName

$w_{allcaps}$	3
$w_{\#ofchars}$	8

player-height

$w_{allcaps}$	-10
$w_{\#ofchars}$	5

player-position

$w_{allcaps}$	3
$w_{\#ofchars}$	4

player-birthName

all caps	True
# of chars	10

player-height

all caps	False
# of chars	3

player-position

all caps	True
# of chars	2

Outline

1 Previous approach: domain-dependent

2 Our approach: domain-independent

3 Similarity features

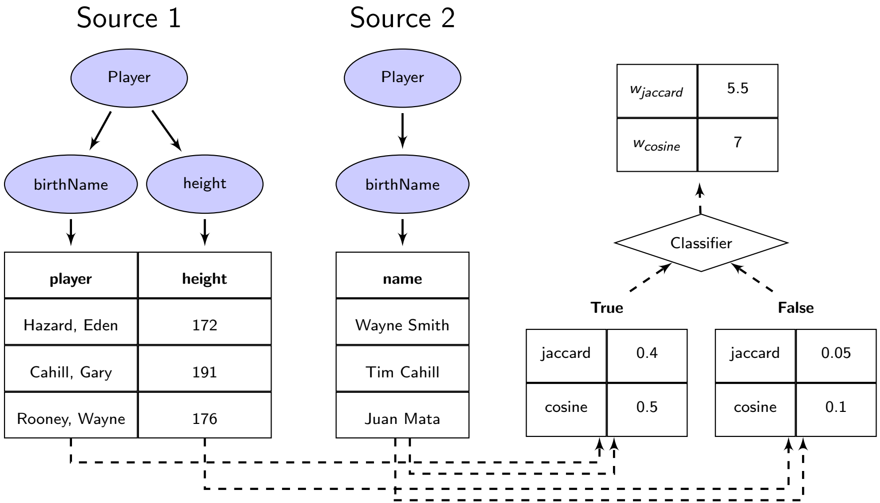
4 Evaluation

5 Conclusion and Future Work

Requirements

- Domain-independent learning models
- Efficient and scalable framework
- Need small amount of domain data as labeled data sources

Our approach: Training



Source 1

player	height
Hazard, Eden	172
Cahill, Gary	191
Rooney, Wayne	176

Source 2

name
Wayne Smith
Tim Cahill
Juan Mata

$w_{jaccard}$	5.5
w_{cosine}	7

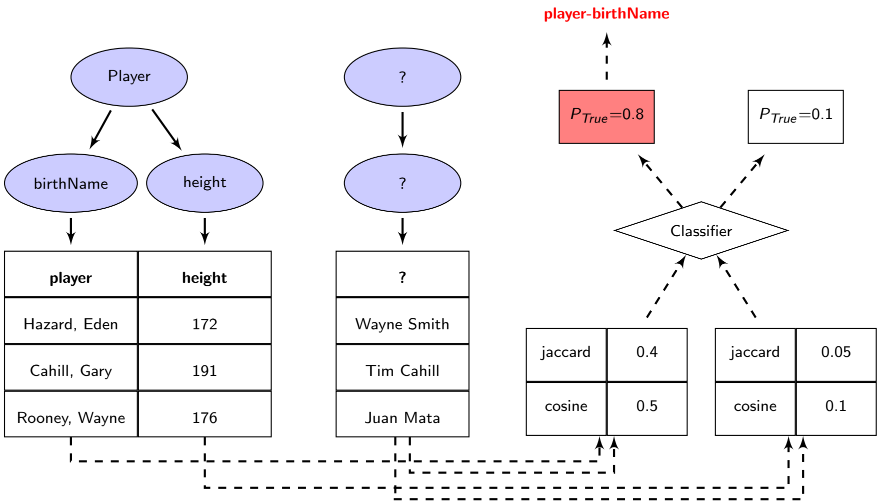
True

jaccard	0.4
cosine	0.5

False

jaccard	0.05
cosine	0.1

Our approach: Predicting



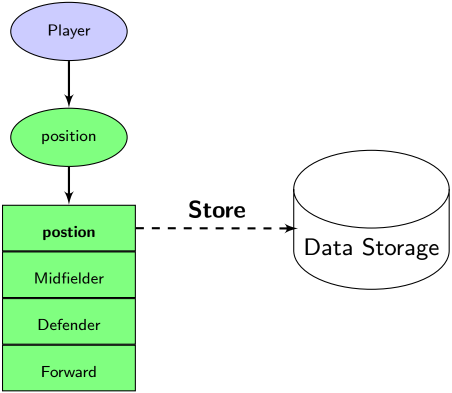
player	height
Hazard, Eden	172
Cahill, Gary	191
Rooney, Wayne	176

?
Wayne Smith
Tim Cahill
Juan Mata

jaccard	0.4
cosine	0.5

jaccard	0.05
cosine	0.1

Our approach: Adding new attribute



Classification models

Classification models:

- Models with class probabilities for ranking scores.
- Typical methods: **Logistic Regression**, Random Forest

Logistic Regression over Random Forest:

- Better interpretation
- Faster training time
- Better class probabilities in ranking situation.

Outline

1 Previous approach: domain-dependent

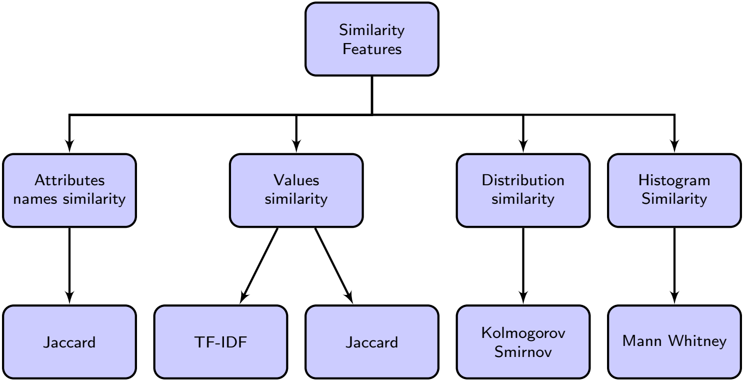
2 Our approach: domain-independent

3 Similarity features

4 Evaluation

5 Conclusion and Future Work

Similarity features



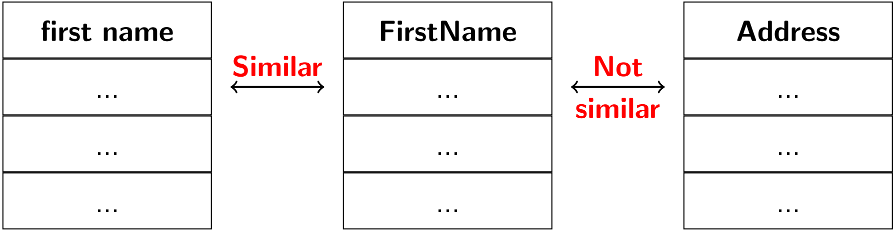
Attribute name similarity

first name
...
...
...

FirstName
...
...
...

Address
...
...
...

Attribute name similarity



first name
...
...
...

Similar

FirstName
...
...
...

Not similar

Address
...
...
...

Similarity measure: Jaccard similarity

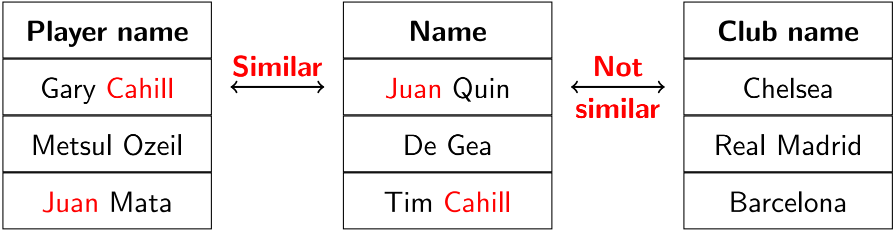
Value similarity

Player name
Gary Cahill
Metsul Ozeil
Juan Mata

Name
Juan Quin
De Gea
Tim Cahill

Club name
Chelsea
Real Madrid
Barcelona

Value similarity



Player name
Gary Cahill
Metsul Ozeil
Juan Mata

Similar

Name
Juan Quin
De Gea
Tim Cahill

Not similar

Club name
Chelsea
Real Madrid
Barcelona

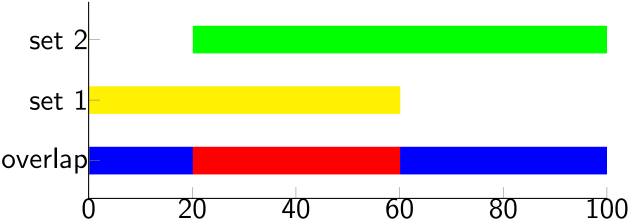
Similarity measures: Jaccard similarity, TF-IDF cosine similarity

Jaccard similarity for numeric values

Numeric Jaccard Similiarity

Given 2 numeric sets of values A, B ranged in $[a_s, a_e]$ and $[b_s, b_e]$:

$$numJaccardSim(A, B) = \frac{|[a_s, a_e] \cap [b_s, b_e]|}{|[a_s, a_e] \cup [b_s, b_e]|}$$

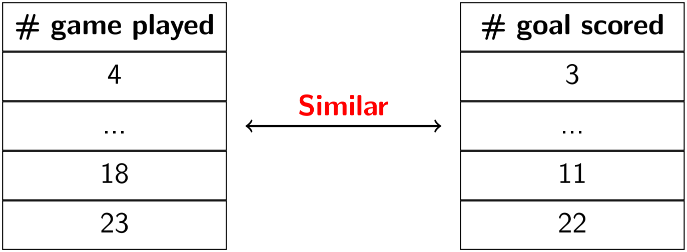


Distribution similarity

game played
4
...
18
23

goal scored
3
...
11
22

Distribution similarity



game played
4
...
18
23

Similar

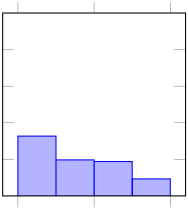
goal scored
3
...
11
22

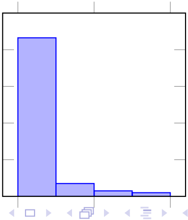
Distribution similarity

game played
4
...
18
23

Similar

goal scored
3
...
11
22





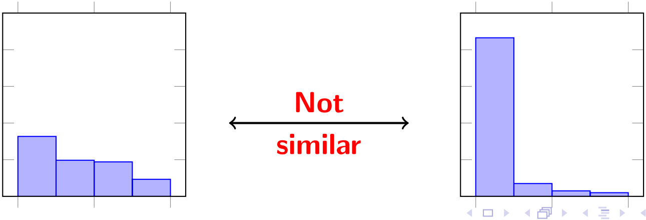
Distribution similarity

game played
4
...
18
23

Similar ⟷

goal scored
3
...
11
22

↓ Similarity measure: KS test ↓

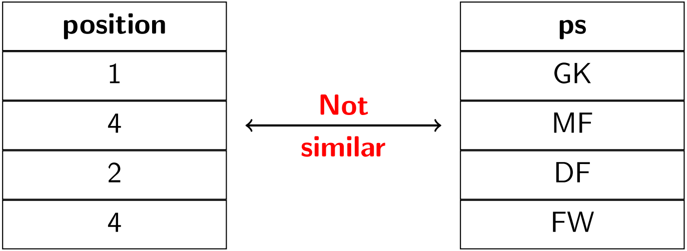


Histogram similarity

position
1
4
2
4

ps
GK
MF
DF
FW

Histogram similarity



position
1
4
2
4

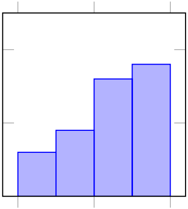
Not similar

ps
GK
MF
DF
FW

Histogram similarity

position
1
4
2
4

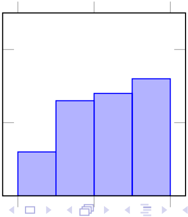
↓



Not similar

ps
GK
MF
DF
FW

↓



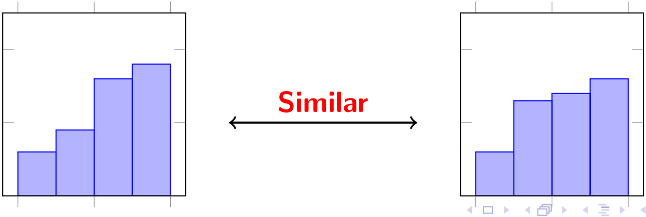
Histogram similarity

position
1
4
2
4

Not similar

ps
GK
MF
DF
FW

↓ Similarity measure: MW test ↓



Outline

1 Previous approach: domain-dependent

2 Our approach: domain-independent

3 Similarity features

4 Evaluation

5 Conclusion and Future Work

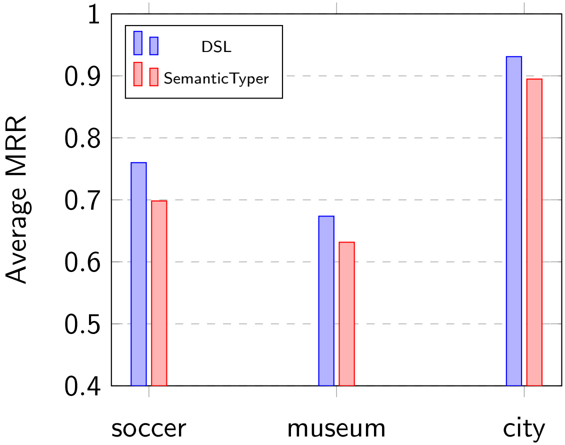
Evaluation

Data sets:

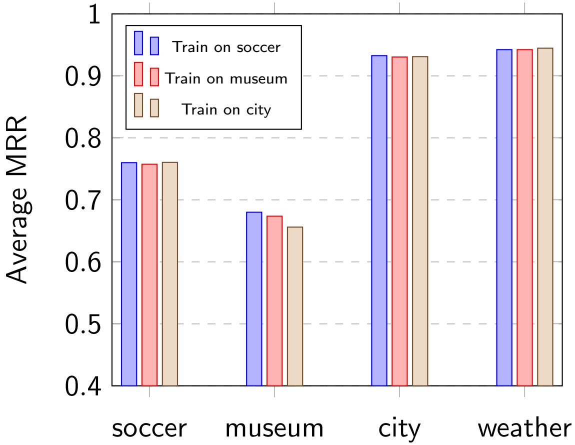
Domain data	# sources	# semantic types	# attributes
soccer	12	14	97
museum	29	20	217
city	10	52	520
weather	4	11	44
T2D Gold	1748	7983	?

Measurements: Mean Reciprocal Rank (MRR)
Evaluating systems: DSL (our approach), SemanticTyper (Ramnandan et al, 2015), T2K (Ritze et al, 2015)

Performance of DSL vs SemanticTyper



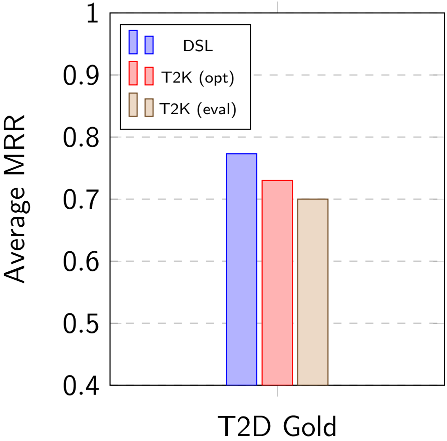
Performance of DSL (trained on different datasets)



Performance of DSL vs T2K on T2D Gold dataset

Experimental settings:

- Labeled sources: DBpedia data in table format
- DSL's classifiers: trained on soccer, museum and city datasets
- T2K results: training and testing



Outline

1 Previous approach: domain-dependent

2 Our approach: domain-independent

3 Similarity features

4 Evaluation

5 Conclusion and Future Work

Conclusion and Future Work

Conclusion:

- Domain-independent approach
- Scalable framework

Future Work:

- Adjust classifier based on domain characteristic
- Detect unseen semantic types in labeling phase