IExM: Information Extraction System for Movies

WWW' 17

Author : Peng-Yu Chen[1], Yi-Hui Lee[2], Yueh-Han Wu[3], Wei-Yun Ma[4]

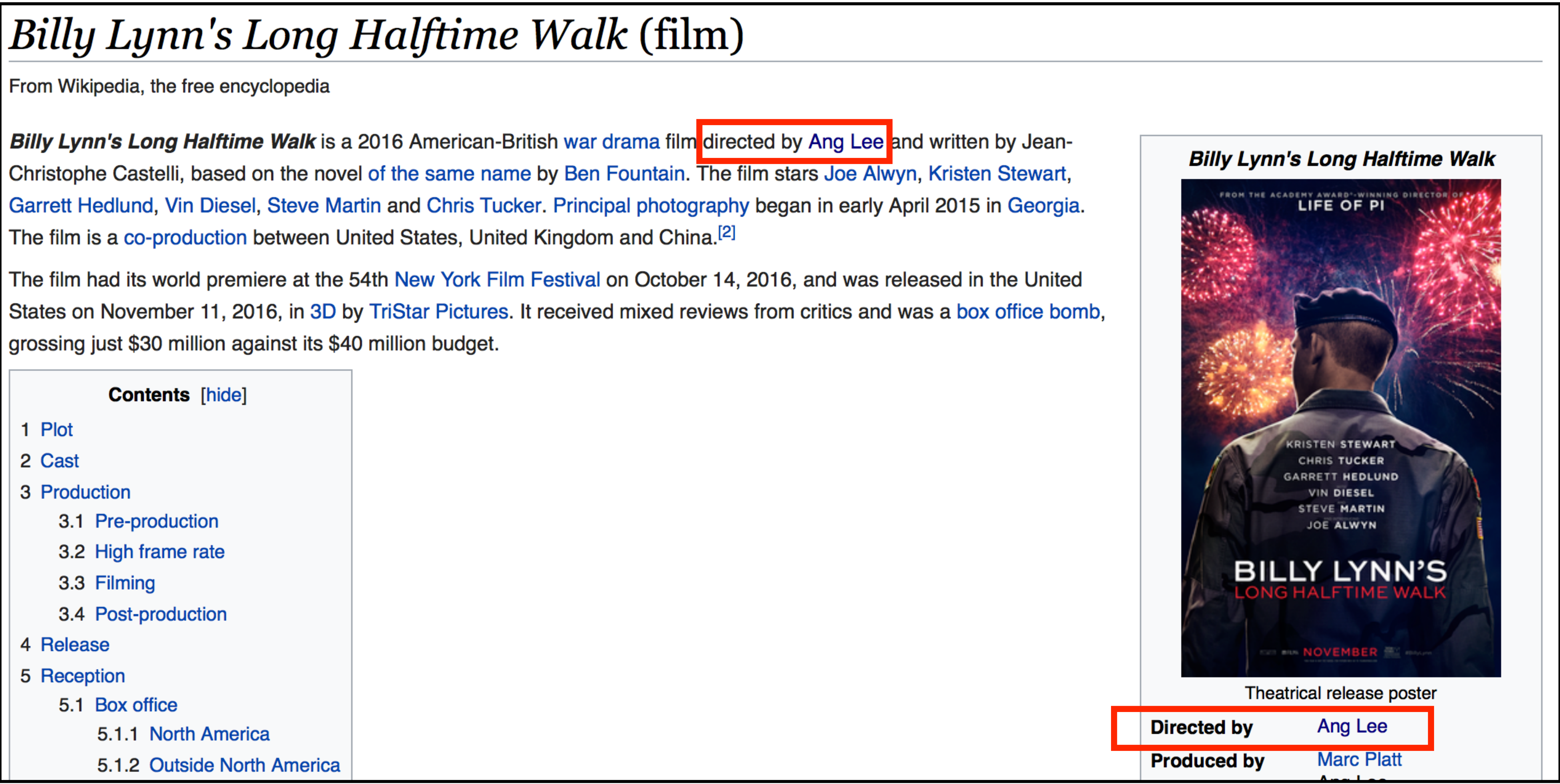
[1, 3]Department of Computer Science National Tsing Hua University,
[2]Department of Computer Science and Information Engineering
National Taiwan Normal University,
[4]Institute of information Science, Academia Sinica
Taiwan, ROC

Outline

- **Introduction**
- Related Work
- Approach
- System: IExM
- Experiment
- Conclusion

Introduction

- Motivation:
  - Wikipedia provides infobox to help users gain the information they want conveniently.
  - Wiki pages with incomplete infobox or without infobox.



Billy Lynn's Long Halftime Walk (film)

From Wikipedia, the free encyclopedia

Billy Lynn's Long Halftime Walk is a 2016 American-British war drama film directed by Ang Lee and written by Jean-Christophe Castelli, based on the novel of the same name by Ben Fountain. The film stars Joe Alwyn, Kristen Stewart, Garrett Hedlund, Vin Diesel, Steve Martin and Chris Tucker. Principal photography began in early April 2015 in Georgia. The film is a co-production between United States, United Kingdom and China.[2]

The film had its world premiere at the 54th New York Film Festival on October 14, 2016, and was released in the United States on November 11, 2016, in 3D by TriStar Pictures. It received mixed reviews from critics and was a box office bomb, grossing just $30 million against its $40 million budget.

Contents [hide]

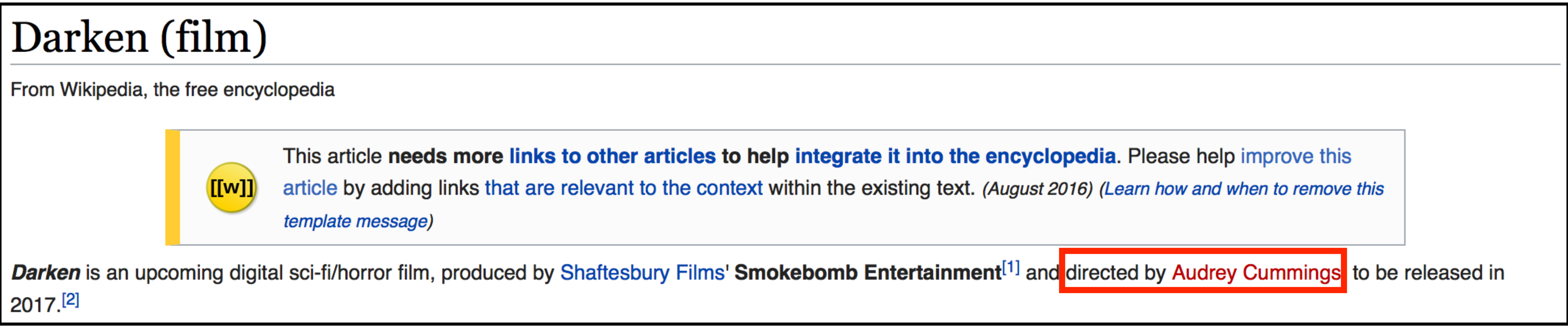
Billy Lynn's Long Halftime Walk

Theatrical release poster

Directed by	Ang Lee
Produced by	Marc Platt

Introduction(cont.)

- Goal:
  - Extract relation instances from unlabeled movie articles

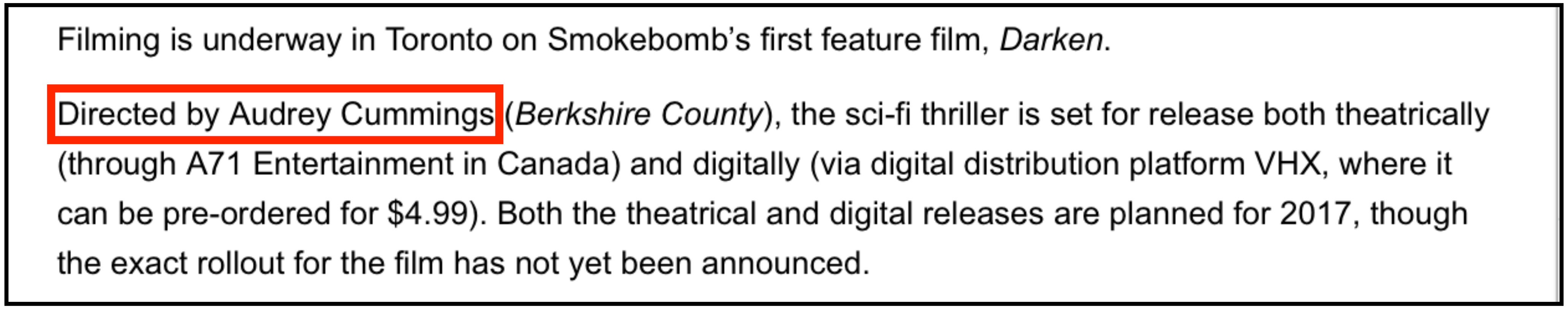


Darken (film)

From Wikipedia, the free encyclopedia

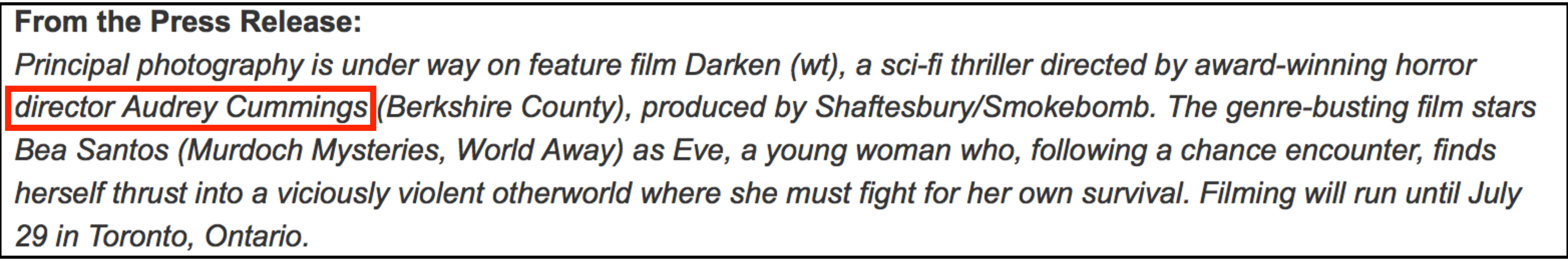
[[w]] This article **needs more links to other articles to help integrate it into the encyclopedia**. Please help improve this article by adding links that are relevant to the context within the existing text. *(August 2016)* *(Learn how and when to remove this template message)*

Darken is an upcoming digital sci-fi/horror film, produced by Shaftesbury Films' **Smokebomb Entertainment**[1] and directed by Audrey Cummings to be released in 2017.[2]



Filming is underway in Toronto on Smokebomb's first feature film, *Darken*.

Directed by Audrey Cummings (*Berkshire County*), the sci-fi thriller is set for release both theatrically (through A71 Entertainment in Canada) and digitally (via digital distribution platform VHX, where it can be pre-ordered for $4.99). Both the theatrical and digital releases are planned for 2017, though the exact rollout for the film has not yet been announced.



From the Press Release:

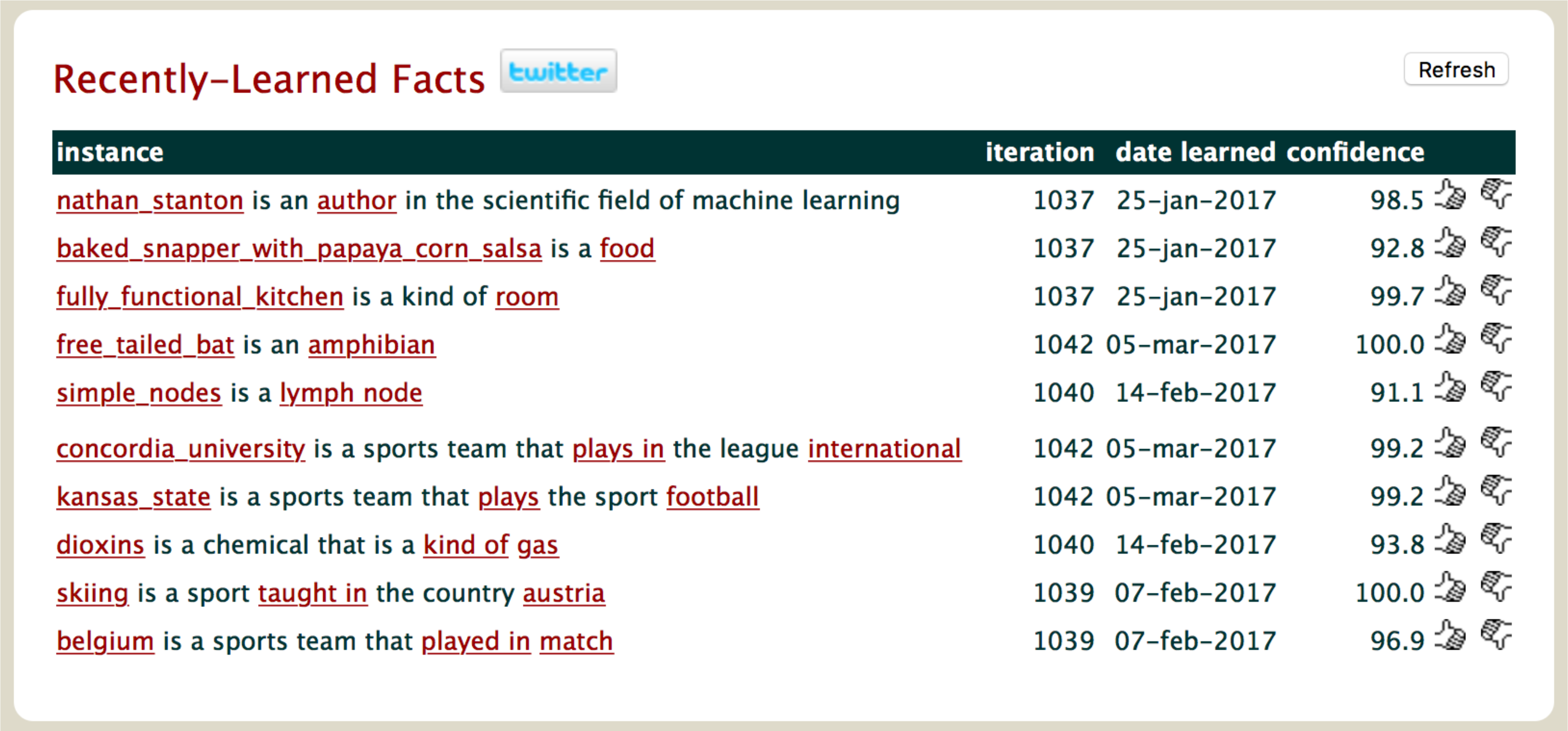
Principal photography is under way on feature film Darken (wt), a sci-fi thriller directed by award-winning horror director Audrey Cummings (Berkshire County), produced by Shaftesbury/Smokebomb. The genre-busting film stars Bea Santos (Murdoch Mysteries, World Away) as Eve, a young woman who, following a chance encounter, finds herself thrust into a viciously violent otherworld where she must fight for her own survival. Filming will run until July 29 in Toronto, Ontario.

Outline

- Introduction
- **Related Work**
- Approach
- System: IExM
- Experiment
- Conclusion

Related Work

- Never-Ending Learning *AAAI' 15*:
  - Read the Web: http://rtw.ml.cmu.edu/rtw/



Recently-Learned Facts

instance	iteration	date learned	confidence	
nathan_stanton is an author in the scientific field of machine learning	1037	25-jan-2017	98.5	👍 👎
baked_snapper_with_papaya_corn_salsa is a food	1037	25-jan-2017	92.8	👍 👎
fully_functional_kitchen is a kind of room	1037	25-jan-2017	99.7	👍 👎
free_tailed_bat is an amphibian	1042	05-mar-2017	100.0	👍 👎
simple_nodes is a lymph node	1040	14-feb-2017	91.1	👍 👎
concordia_university is a sports team that plays in the league international	1042	05-mar-2017	99.2	👍 👎
kansas_state is a sports team that plays the sport football	1042	05-mar-2017	99.2	👍 👎
dioxins is a chemical that is a kind of gas	1040	14-feb-2017	93.8	👍 👎
skiing is a sport taught in the country austria	1039	07-feb-2017	100.0	👍 👎
belgium is a sports team that played in match	1039	07-feb-2017	96.9	👍 👎

Related Work(cont.)

- Never-Ending Learning *AAAI’ 15*:
  - Contribution:

    - Couple training

      eg. **serverdWith(tea, biscuits)**

    - Semi-supervised learning pattern

    - Mutual exclusive constraint strategy
  - Weakness:

    - Without pattern ranking strategy

Related Work(cont.)

- Semi-supervised Semantic Pattern Discovery with Guidance from Unsupervised Pattern Clusters *Coling' 10*:
  - Contribution:

    - Pattern ranking algorithm

    - Prevent semantic drift
  - Weakness:

    - Accept top ranked patterns only

    - Does not update patterns' qualities that patterns actually generated.

    - Estimates patterns' quality only based on the instances (and their clusters) that these patterns can match.

Outline

- Introduction
- Related Work
- **Approach**
- System: IExM
- Experiment
- Conclusion

Approach

- Our **I**mproved **P**attern **R**anking **A**lgorithm (**IPRA**):

  - Extract attributes

  - Can't use MutualExclusive as constraints cause our topic only focus on movie

  - Estimates patterns' quality according to various factors:

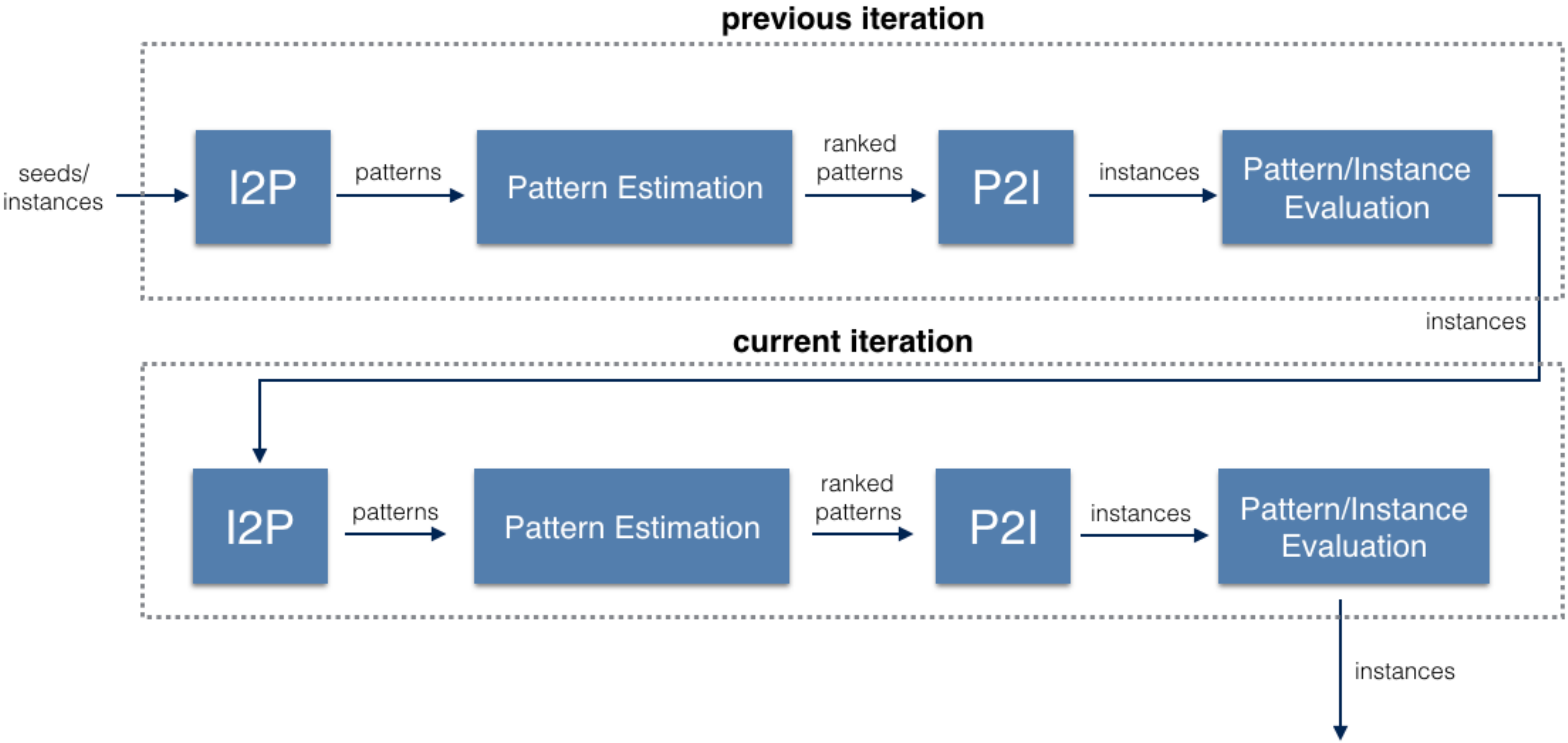
    - Occurrence of application

    - Coverage of application

    - The quality estimation of the instances which are actually extracted by these patterns

Approach(cont.)

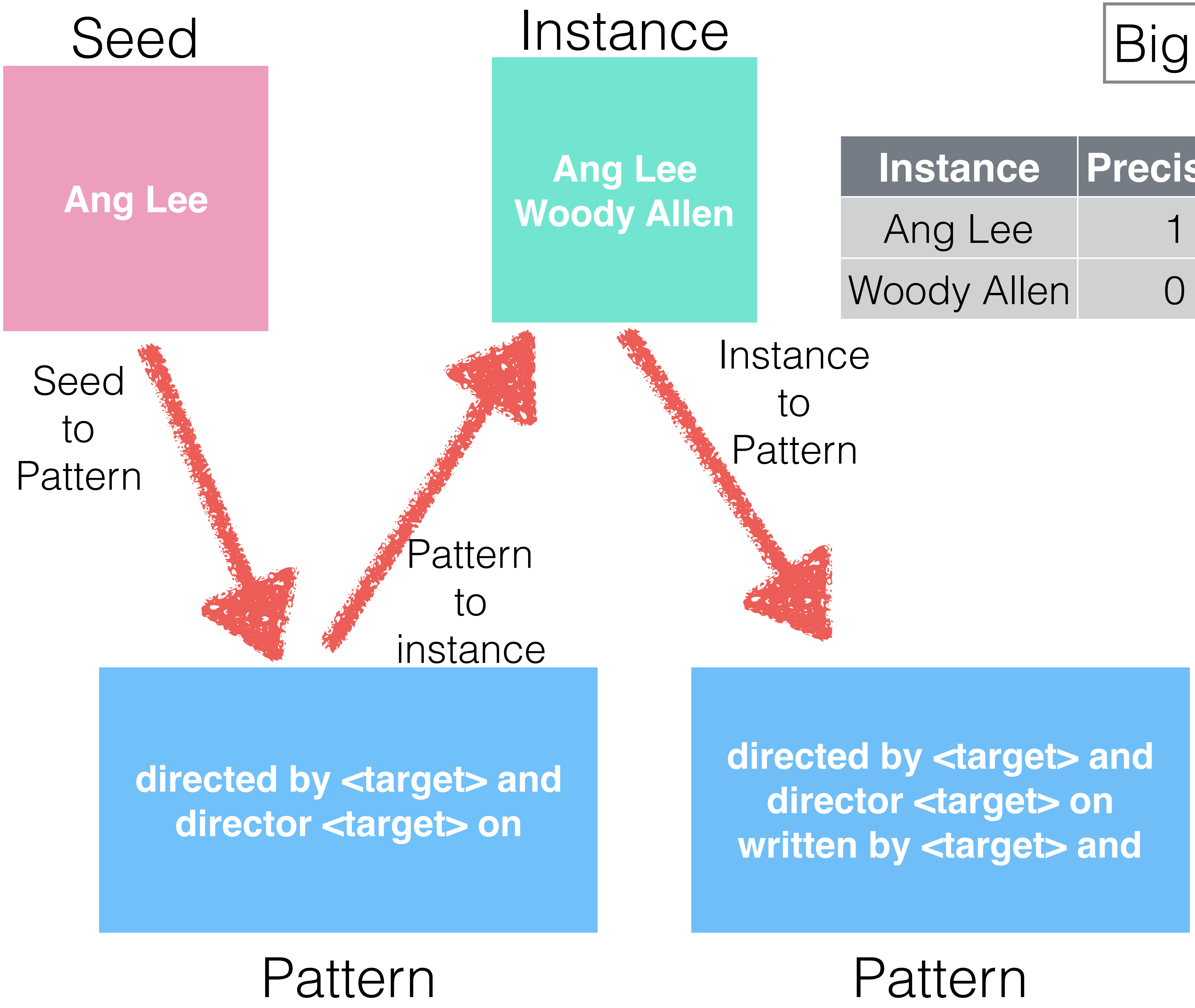
- **IPRA** Framework:



Approach(cont.)

Big picture

Instance	Precision
Ang Lee	1
Woody Allen	0



Seed

Ang Lee

Instance

Ang Lee
Woody Allen

Seed to Pattern

Pattern to instance

Instance to Pattern

directed by <target> and director <target> on

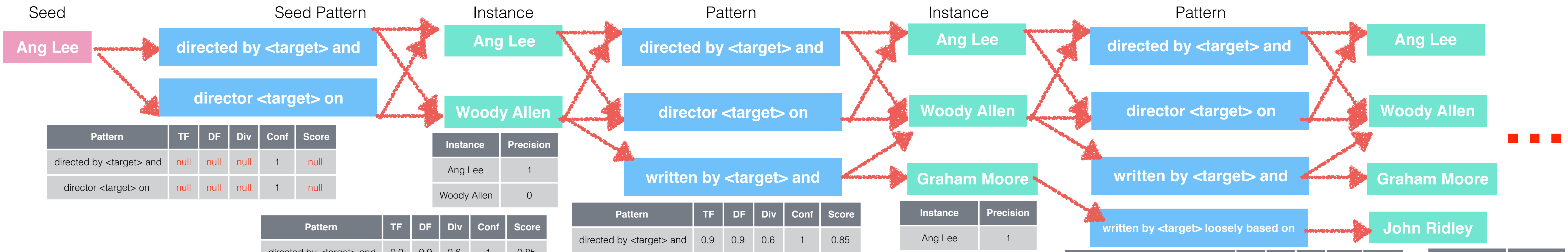
Pattern

directed by <target> and director <target> on written by <target> and

Pattern

Details

Seed to Pattern | Pattern to instance | Instance to Pattern | Pattern to instance | Instance to Pattern



Pattern	TF	DF	Div	Conf	Score
directed by <target> and	null	null	null	1	null
director <target> on	null	null	null	1	null

Pattern	TF	DF	Div	Conf	Score
directed by <target> and	0.9	0.9	0.6	1	0.85
director <target> on	0.8	0.8	0.7	1	0.825

$$InstanceScore(I_i) = \frac{\sum_{j=1}^{k}(PatternScore(P_j) \times \frac{1}{rank(P_j)})}{\sum_{j=1}^{k}\frac{1}{rank(P_j)}}$$

$$EstimatedPatternScore(P_i) = \frac{\sum_{j=1}^{k} InstanceScore(I_j)}{k}$$

(0.85*1+0.825*(1/2))/(1+(1/2))=0.842

Instance	Precision
Ang Lee	1
Woody Allen	0

Pattern	TF	DF	Div	Conf	Score
directed by <target> and	0.9	0.9	0.6	1	0.85
director <target> on	0.8	0.8	0.7	1	0.825
written by <target> and	null	null	null	0	null

$$Conf(P_i) = 1 - \prod_{j=1}^{k}(1 - Prec(I_j))$$

Pattern	TF	DF	Div	Conf	Score
directed by <target> and	0.9	0.9	0.6	1	0.85
director <target> on	0.8	0.8	0.7	1	0.825
written by <target> and	null	null	null	0	0.842

Instance	Precision
Ang Lee	1
Woody Allen	2/3
Graham Moore	0

$$Prec(I_i) = \frac{\sum_{j=1}^{k} Conf(P_j)}{k}$$

Pattern	TF	DF	Div	Conf	Score
directed by <target> and	0.9	0.9	0.6	1	0.85
director <target> on	0.8	0.8	0.7	1	0.825
written by <target> and	0.8	0.8	0.8	0	0.8

Pattern	TF	DF	Div	Conf	Score
directed by <target> and	0.9	0.9	0.6	1	0.85
director <target> on	0.8	0.8	0.7	1	0.825
written by <target> and	0.8	0.8	0.8	0	0.8
written by <target> loosely based on	null	null	null	0	null

(0.85*1+0.825*(1/2))/(1+(1/2))=0.842

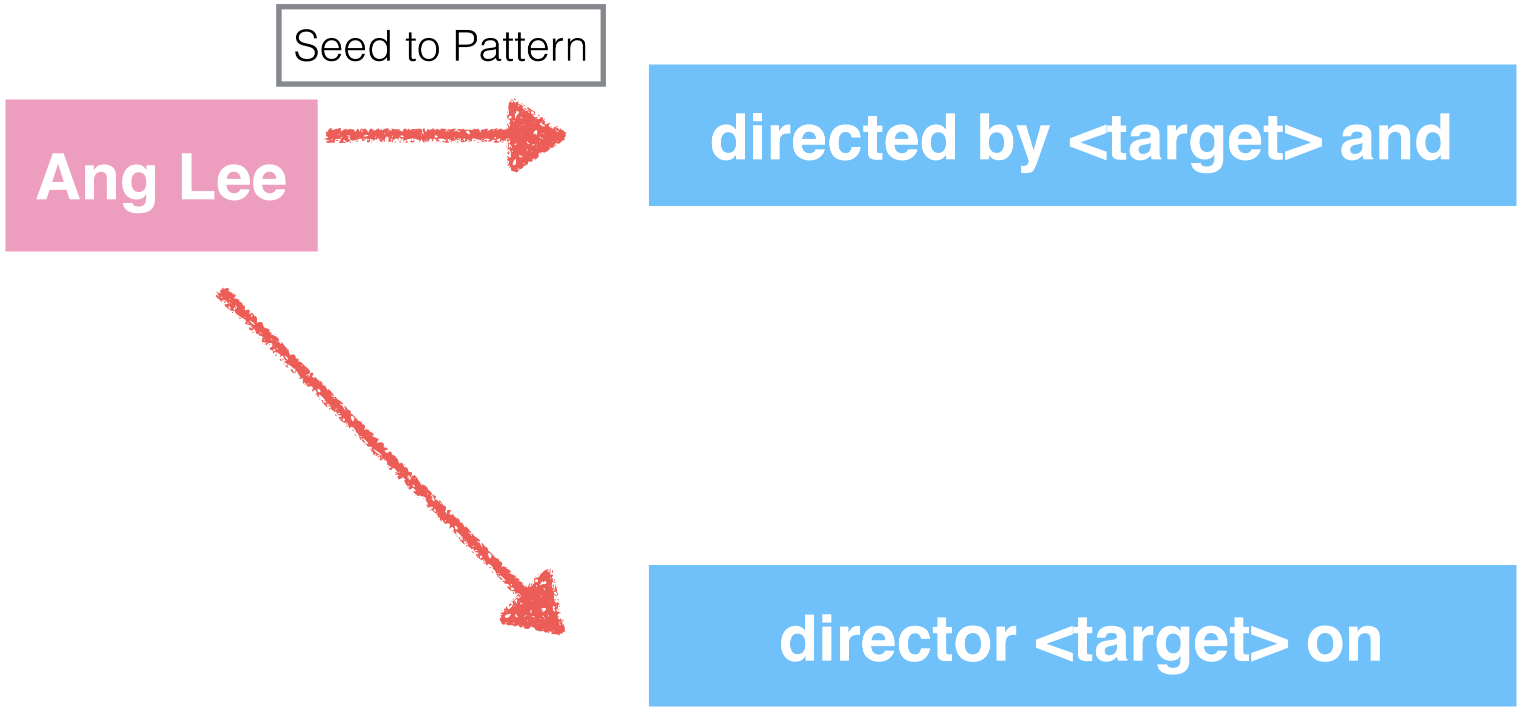
Pattern	TF	DF	Div	Conf	Score
directed by <target> and	0.9	0.9	0.6	1	0.85
director <target> on	0.8	0.8	0.7	1	0.825
written by <target> and	0.8	0.8	0.8	0	0.8
written by <target> loosely based on	null	null	null	0	0.8

Instance	Precision
Ang Lee	1
Woody Allen	2/3
Graham Moore	0
John Ridley	0

Seed

Seed Pattern

Details



Seed to Pattern

Ang Lee

directed by <target> and

director <target> on

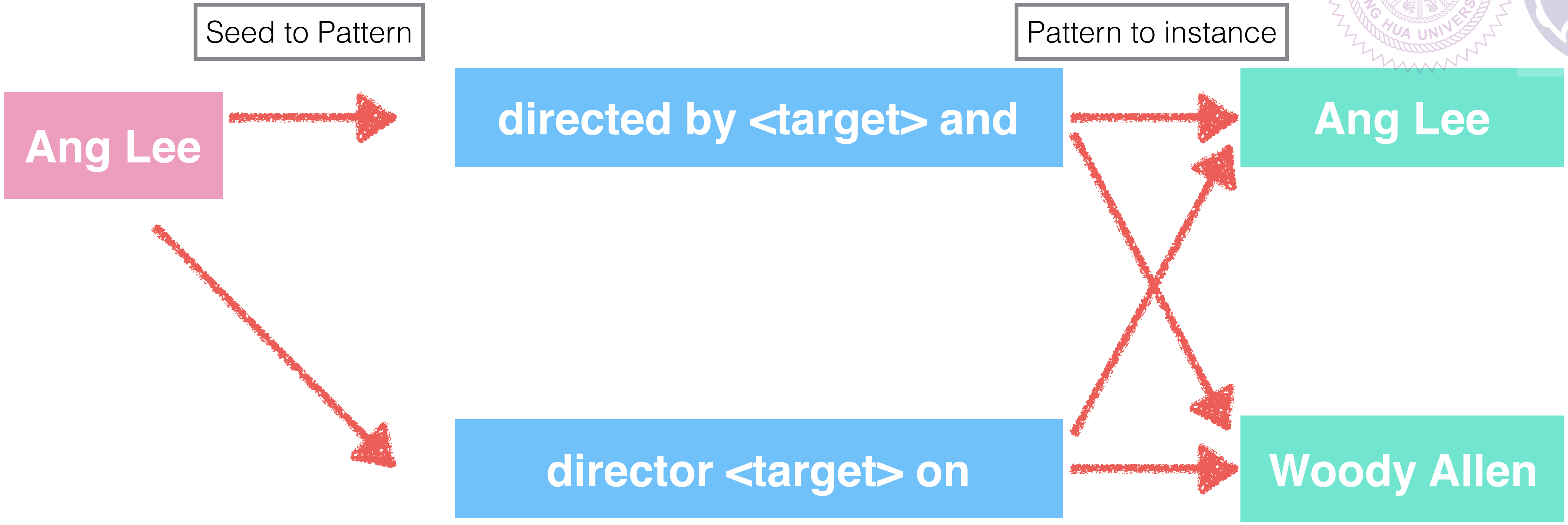
Pattern	TF	DF	Div	Conf	Score
directed by <target> and	null	null	null	1	null
director <target> on	null	null	null	1	null

Seed

Seed Pattern

Instance

Details



Seed to Pattern

Pattern to instance

Ang Lee

directed by <target> and

director <target> on

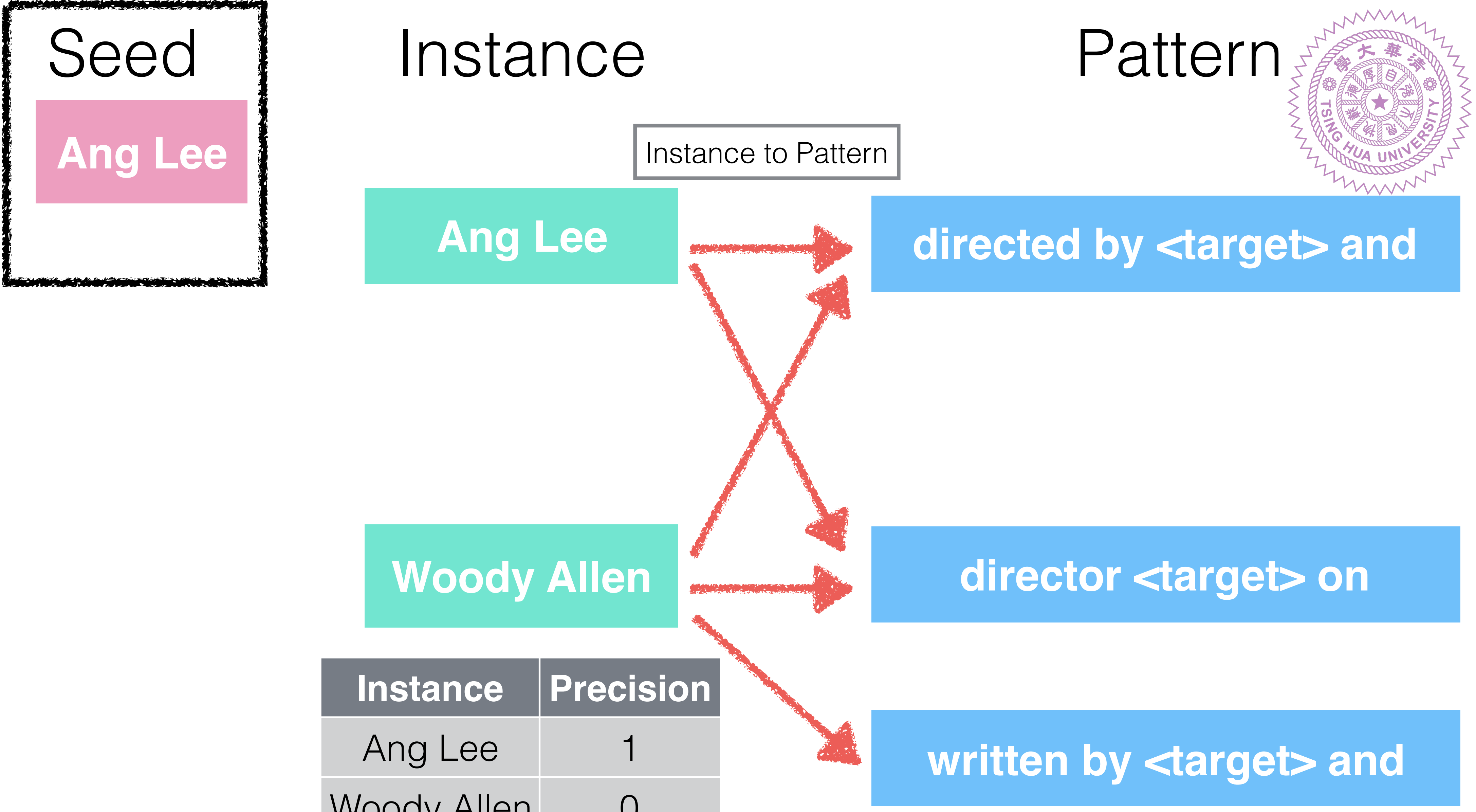
Ang Lee

Woody Allen

Pattern	TF	DF	Div	Conf	Score
directed by <target> and	null	null	null	1	null
director <target> on	null	null	null	1	null

Instance	Precision
Ang Lee	1
Woody Allen	0

Pattern	TF	DF	Div	Conf	Score
directed by <target> and	0.9	0.9	0.6	1	0.85
director <target> on	0.8	0.8	0.7	1	0.825



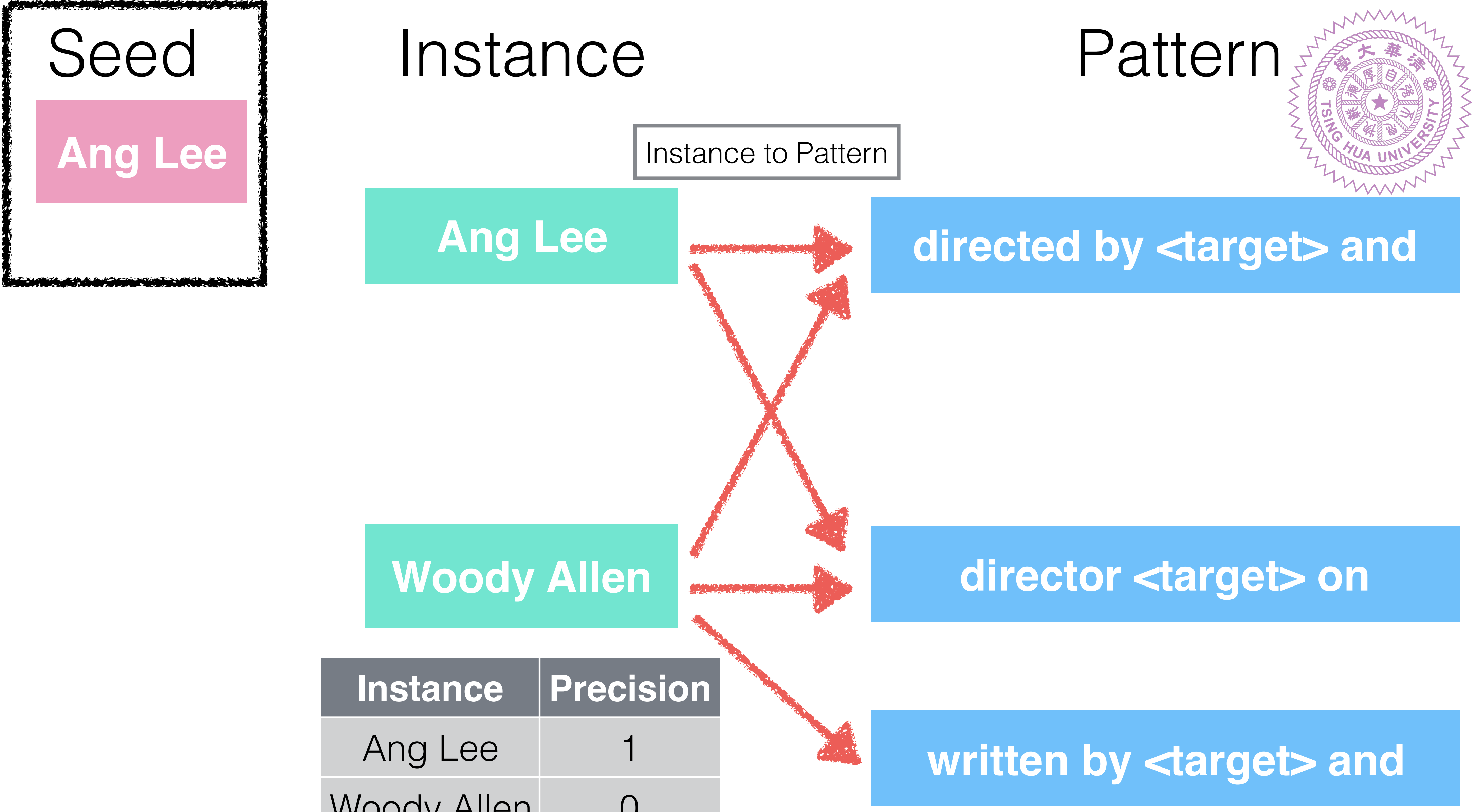
Details

Instance	Precision
Ang Lee	1
Woody Allen	0

Pattern	TF	DF	Div	Conf	Score
directed by <target> and	0.9	0.9	0.6	1	0.85
director <target> on	0.8	0.8	0.7	1	0.825

Pattern	TF	DF	Div	Conf	Score
directed by <target> and	0.9	0.9	0.6	1	0.85
director <target> on	0.8	0.8	0.7	1	0.825
written by <target> and	null	null	null	**0**	null

$$Conf(P_i) = 1 - \prod_{j=1}^{k} (1 - Prec(I_j))$$



Details

Instance	Precision
Ang Lee	1
Woody Allen	0

Pattern	TF	DF	Div	Conf	Score
directed by <target> and	0.9	0.9	0.6	1	0.85
director <target> on	0.8	0.8	0.7	1	0.825

Pattern	TF	DF	Div	Conf	Score
directed by <target> and	0.9	0.9	0.6	1	0.85
director <target> on	0.8	0.8	0.7	1	0.825
written by <target> and	null	null	null	0	**0.842**

$$InstanceScore(I_i) = \frac{\sum_{j=1}^{k}(PatternScore(P_j) \times \frac{1}{rank(P_j)})}{\sum_{j=1}^{k}\frac{1}{rank(P_j)}}$$

↓

$$EstimatedPatternScore(P_i) = \frac{\sum_{j=1}^{k} InstanceScore(I_j)}{k}$$

(0.85*1+0.825*(1/2))/(1+(1/2))=0.842

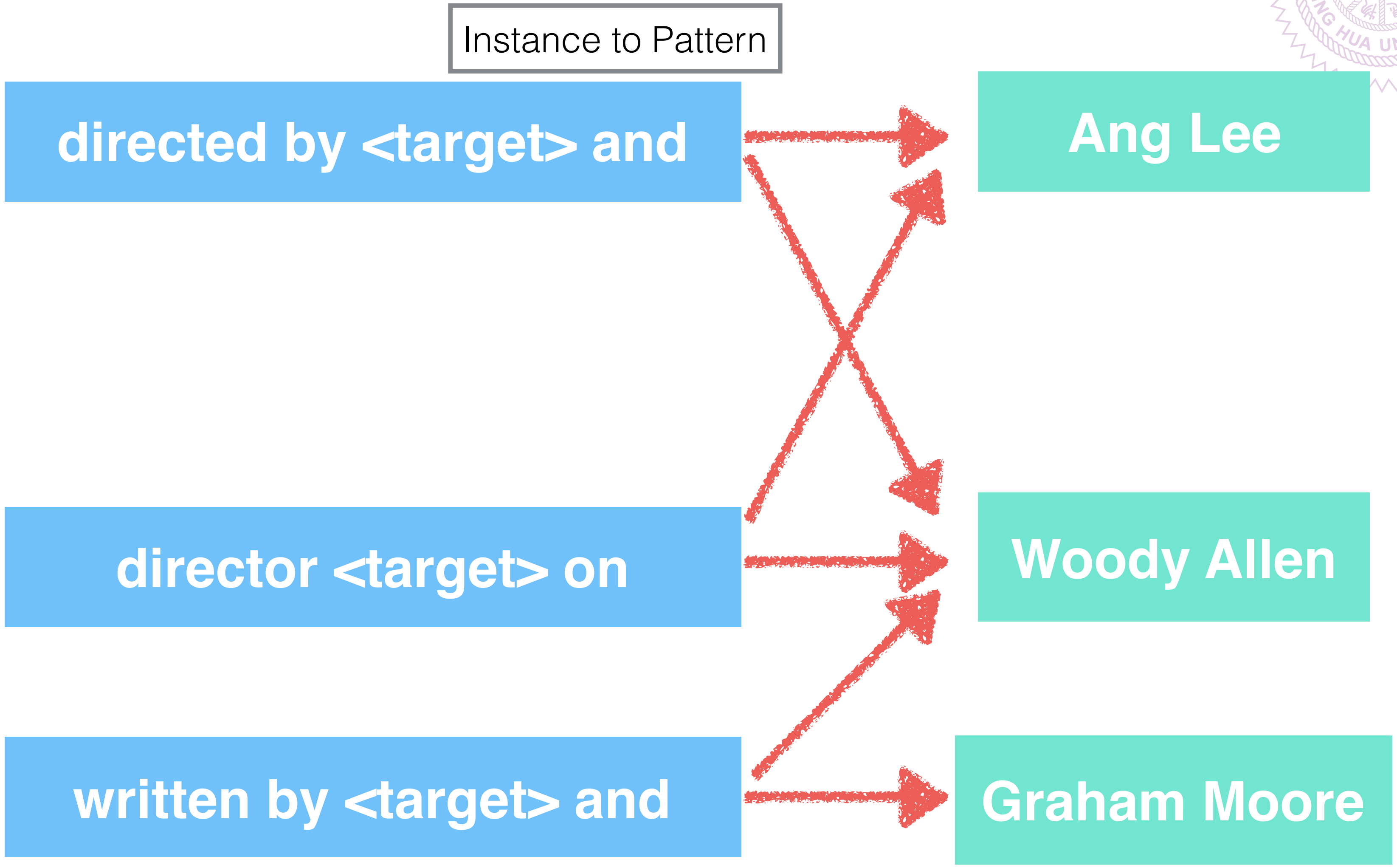
Seed

Ang Lee

Pattern

Instance

Details



$$Prec(I_i) = \frac{\sum_{j=1}^{k} Conf(P_j)}{k}$$

Pattern	TF	DF	Div	Conf	Score
directed by <target> and	0.9	0.9	0.6	1	0.85
director <target> on	0.8	0.8	0.7	1	0.825
written by <target> and	null	null	null	0	0.842

Instance	Precision
Ang Lee	1
Woody Allen	2/3
Graham Moore	0

Pattern	TF	DF	Div	Conf	Score
directed by <target> and	0.9	0.9	0.6	1	0.85
director <target> on	0.8	0.8	0.7	1	0.825
written by <target> and	0.8	0.8	0.8	0	0.8

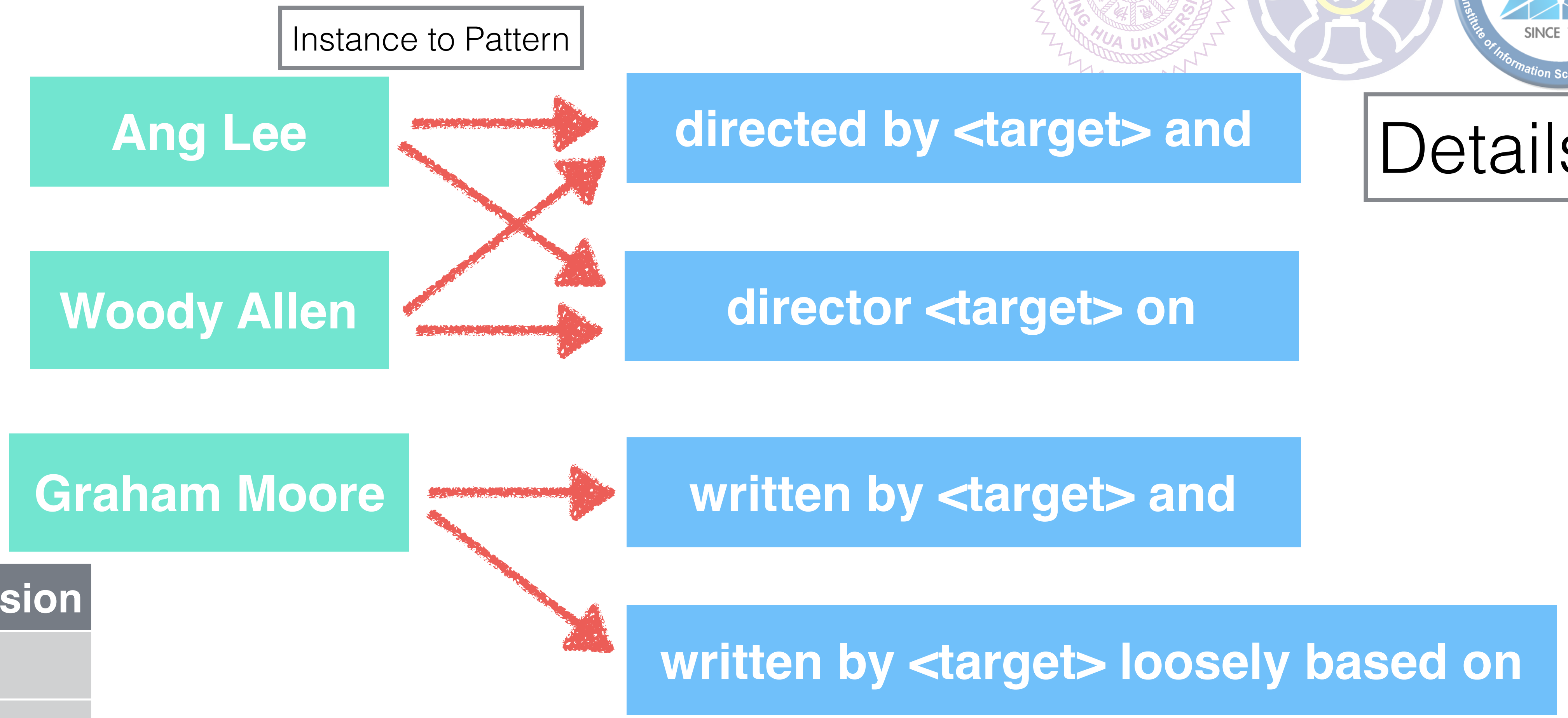
Seed

Ang Lee

Instance

Pattern

Details



Instance to Pattern

Ang Lee

Woody Allen

Graham Moore

directed by <target> and

director <target> on

written by <target> and

written by <target> loosely based on

Instance	Precision
Ang Lee	1
Woody Allen	2/3
Graham Moore	0

Pattern	TF	DF	Div	Conf	Score
directed by <target> and	0.9	0.9	0.6	1	0.85
director <target> on	0.8	0.8	0.7	1	0.825
written by <target> and	0.8	0.8	0.8	0	0.8

Pattern	TF	DF	Div	Conf	Score
directed by <target> and	0.9	0.9	0.6	1	0.85
director <target> on	0.8	0.8	0.7	1	0.825
written by <target> and	0.8	0.8	0.8	0	0.8
written by <target> loosely based on	null	null	null	0	null

Pattern	TF	DF	Div	Conf	Score
directed by <target> and	0.9	0.9	0.6	1	0.85
director <target> on	0.8	0.8	0.7	1	0.825
written by <target> and	0.8	0.8	0.8	0	0.8
written by <target> loosely based on	null	null	null	0	**0.8**

$$InstanceScore(I_i) = \frac{\sum_{j=1}^{k}(PatternScore(P_j) \times \frac{1}{rank(P_j)})}{\sum_{j=1}^{k}\frac{1}{rank(P_j)}}$$

$$EstimatedPatternScore(P_i) = \frac{\sum_{j=1}^{k} InstanceScore(I_j)}{k}$$

0.8*(1/3)/(1/3) = 0.8

Seed

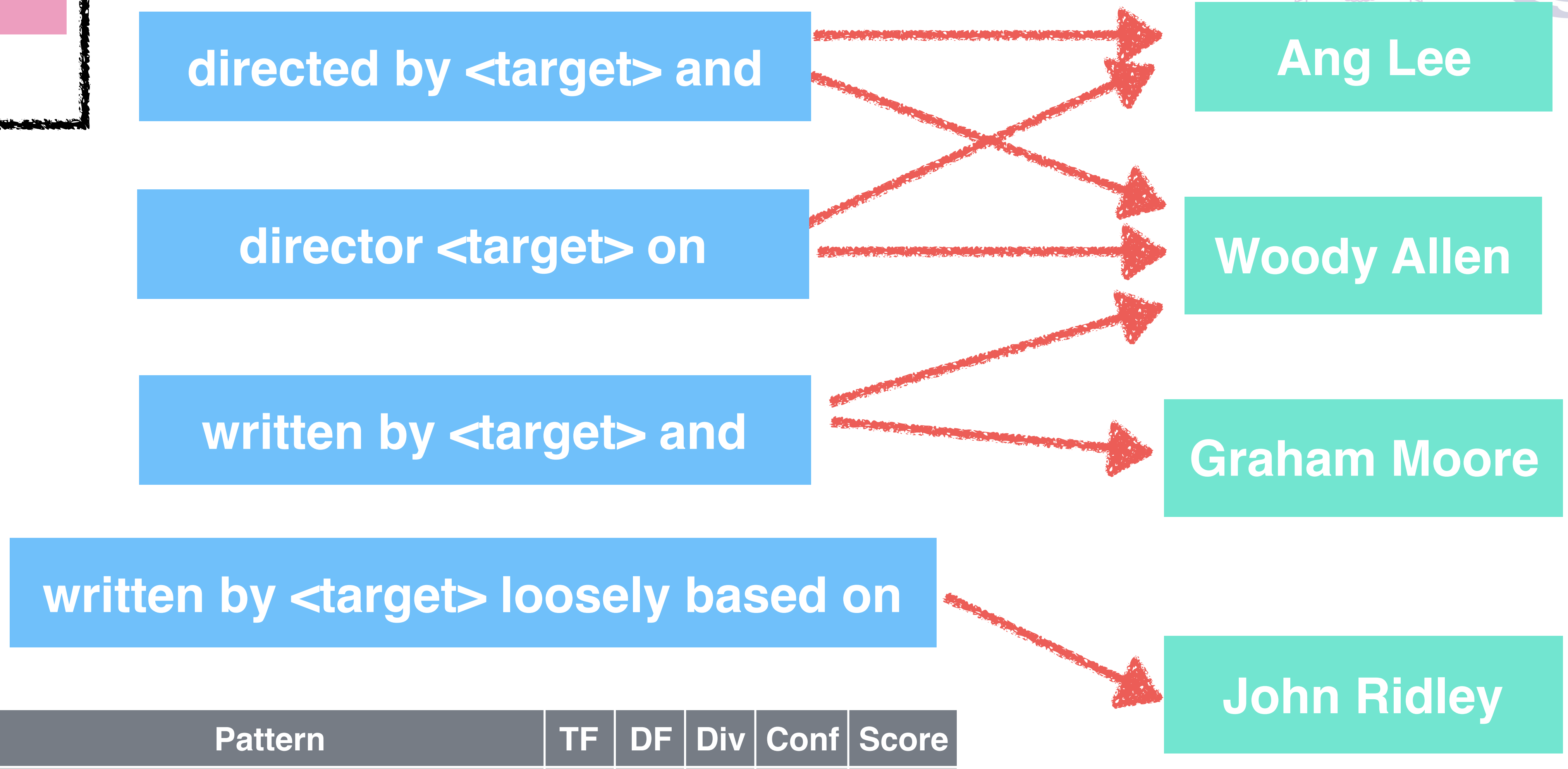
Ang Lee

Pattern

Instance

Instance to Pattern

Details



directed by <target> and

director <target> on

written by <target> and

written by <target> loosely based on

Ang Lee

Woody Allen

Graham Moore

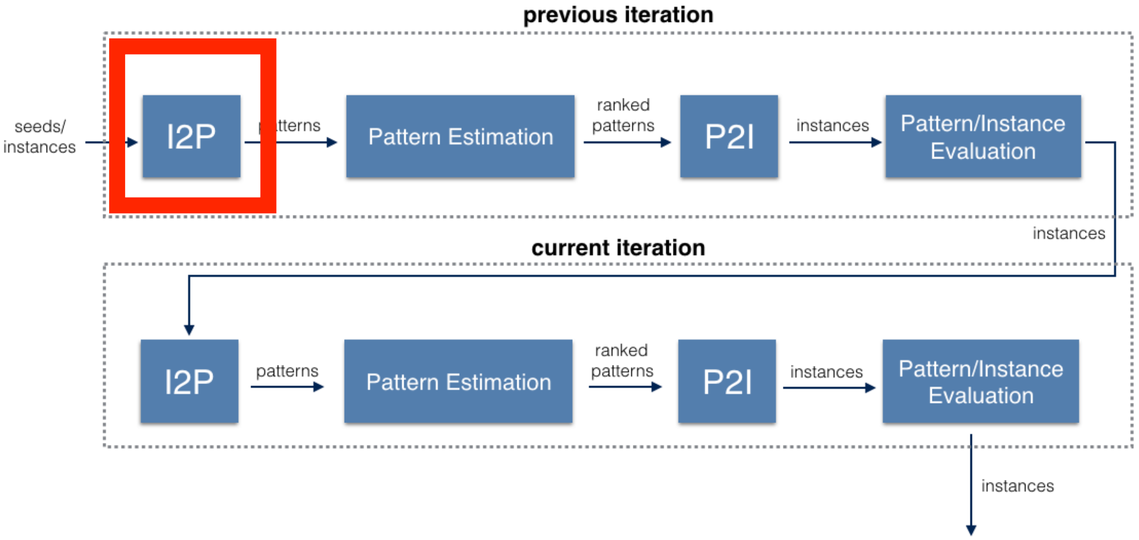
John Ridley

...

Pattern	TF	DF	Div	Conf	Score
directed by <target> and	0.9	0.9	0.6	1	0.85
director <target> on	0.8	0.8	0.7	1	0.825
written by <target> and	0.8	0.8	0.8	0	0.8
written by <target> loosely based on	null	null	null	0	null

Instance	Precision
Ang Lee	1
Woody Allen	2/3
Graham Moore	0
John Ridley	0

Pattern	TF	DF	Div	Conf	Score
directed by <target> and	0.9	0.9	0.6	1	0.85
director <target> on	0.8	0.8	0.7	1	0.825
written by <target> and	0.8	0.8	0.8	0	0.8
written by <target> loosely based on	null	null	null	0	0.8



Approach(cont.)

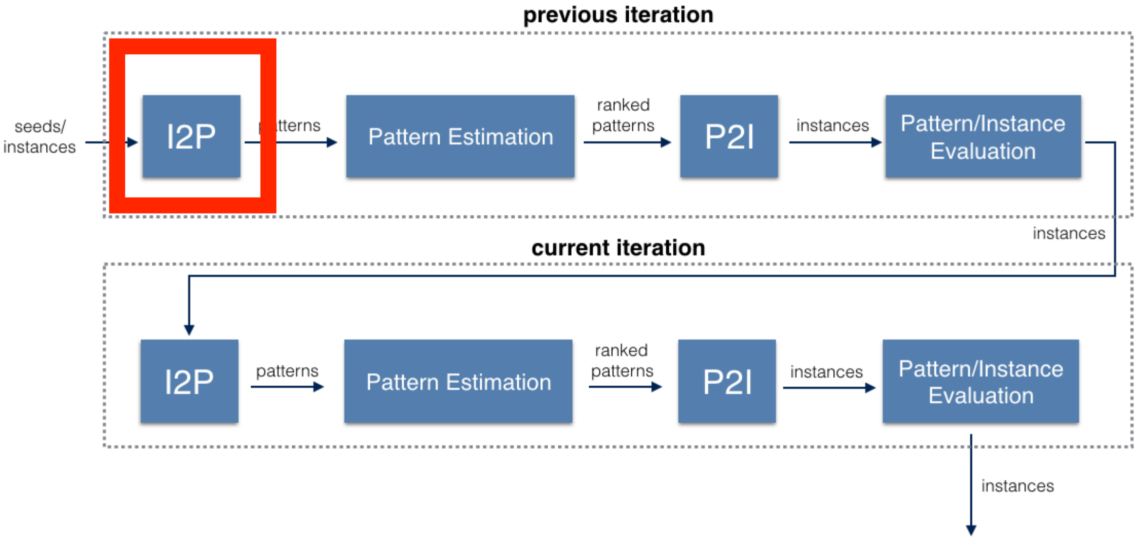
- Pattern Design:

Context

Word-only
POS
E-HowNet Wordsense

Syntactic

Parse path
Parse path + head



Approach(cont.)

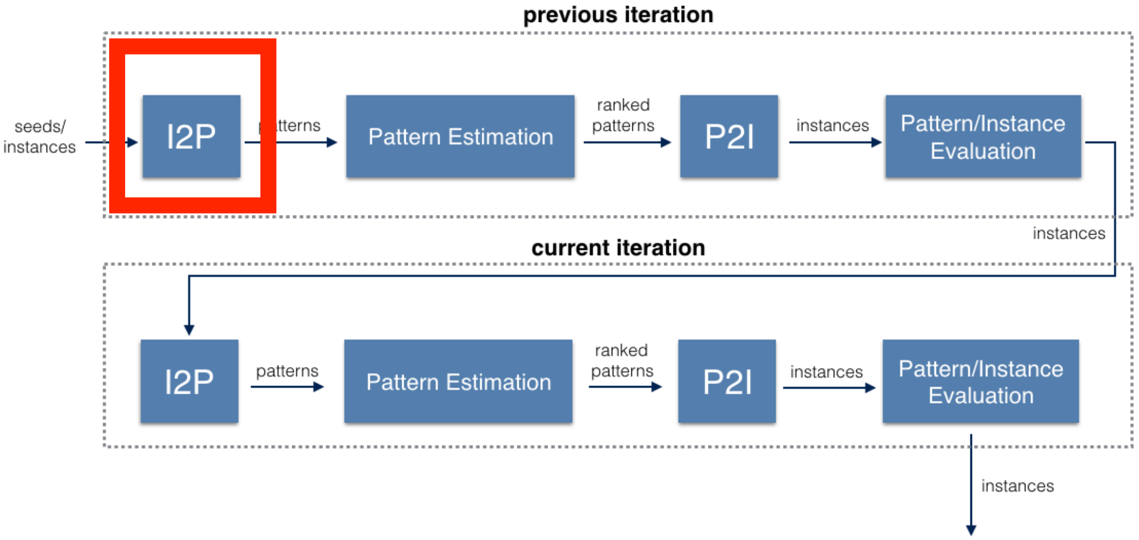
- Pattern Design: window size = 1

Context

Content: Billy Lynn's Long Halftime Walk is a 2016 American-British war drama film directed by Ang Lee and written by Jean-Christophe Castelli.

Tokenize and Tagged: Billy(NNP) Lynn(NNP) 's(POS) Long(NNP) Halftime(NNP) Walk(NNP) is(VBZ) a(DT) 2016(CD) American-British(JJ) war(NN) drama(NN) film(NN) directed(VBN) by(IN) Ang(NNP) Lee(NNP) and(CC) written(VBN) by(IN) Jean-Christophe(NNP) Castelli(NNP) .(.)

target



Approach(cont.)

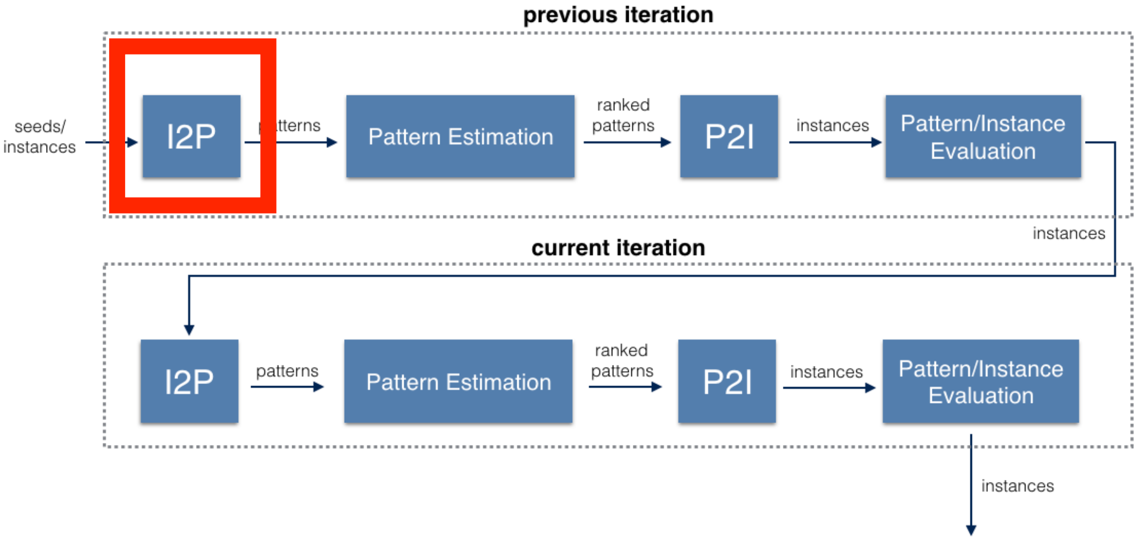
- Pattern Design:

Context

Word-only
POS
E-HowNet Wordsense

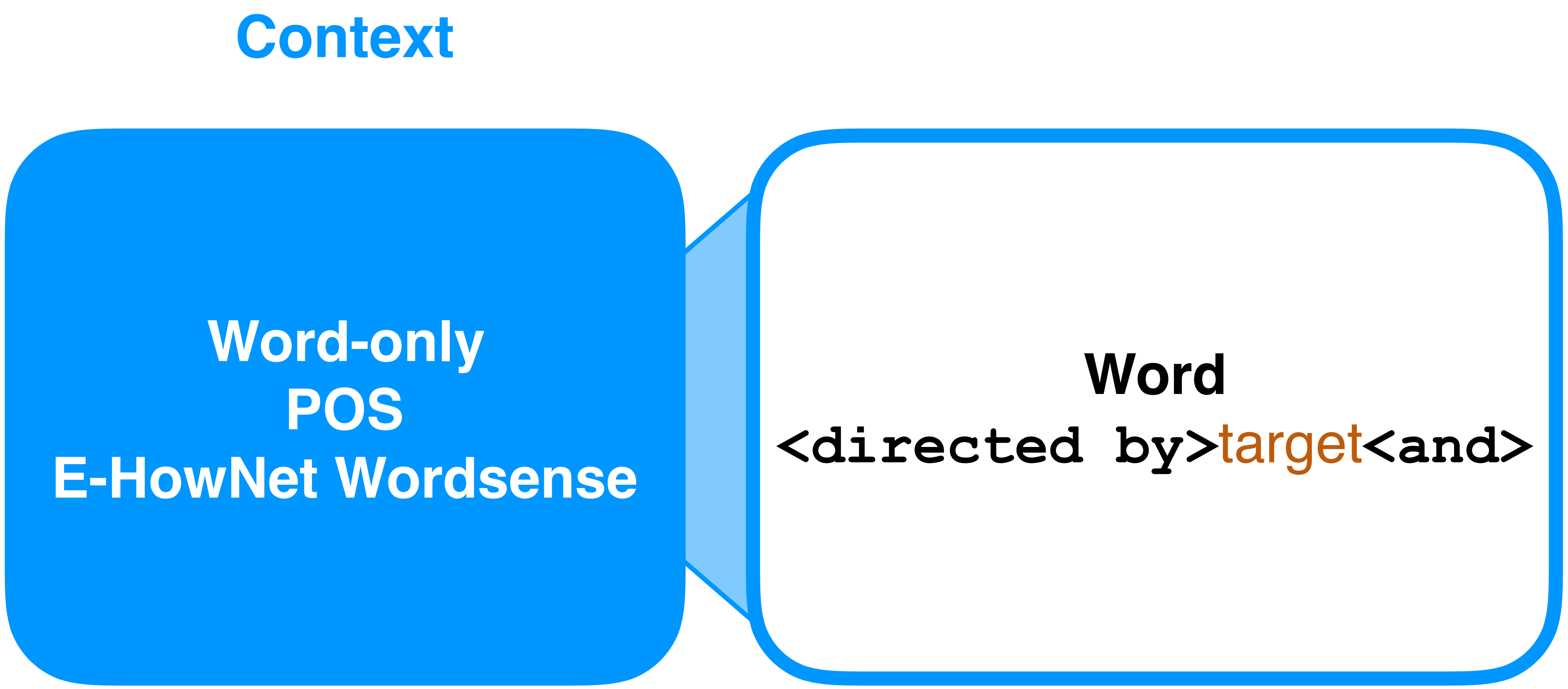
Syntactic

Parse path
Parse path + head



Approach(cont.)

- Pattern Design:

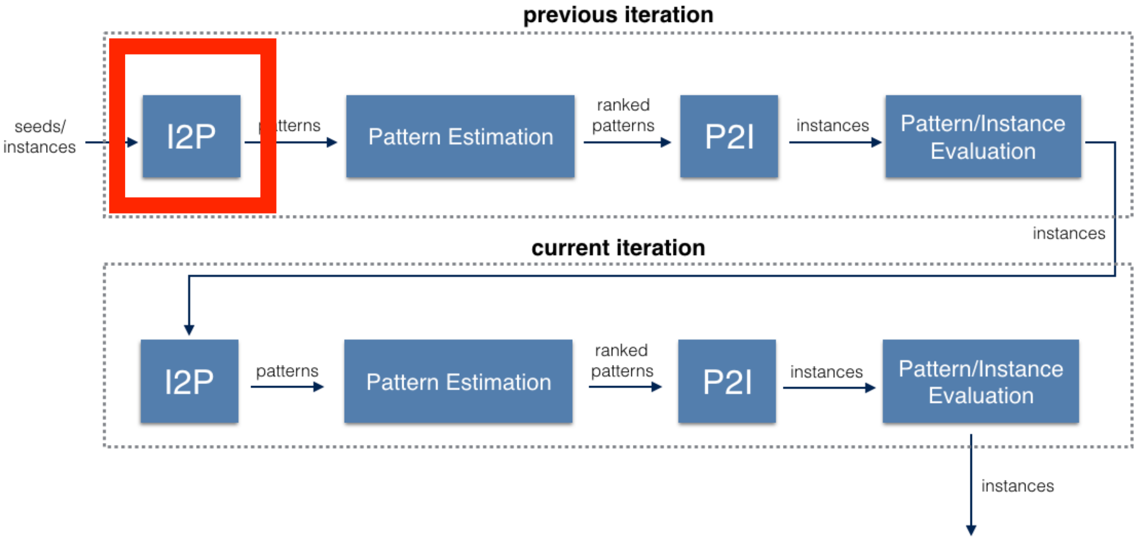


Context

Word-only
POS
E-HowNet Wordsense

Word
`<directed by>`target**`<and>`**

Approach(cont.)


previous iteration

seeds/ instances

I2P

patterns

Pattern Estimation

ranked patterns

P2I

instances

Pattern/Instance Evaluation

instances

current iteration

I2P

patterns

Pattern Estimation

ranked patterns

P2I

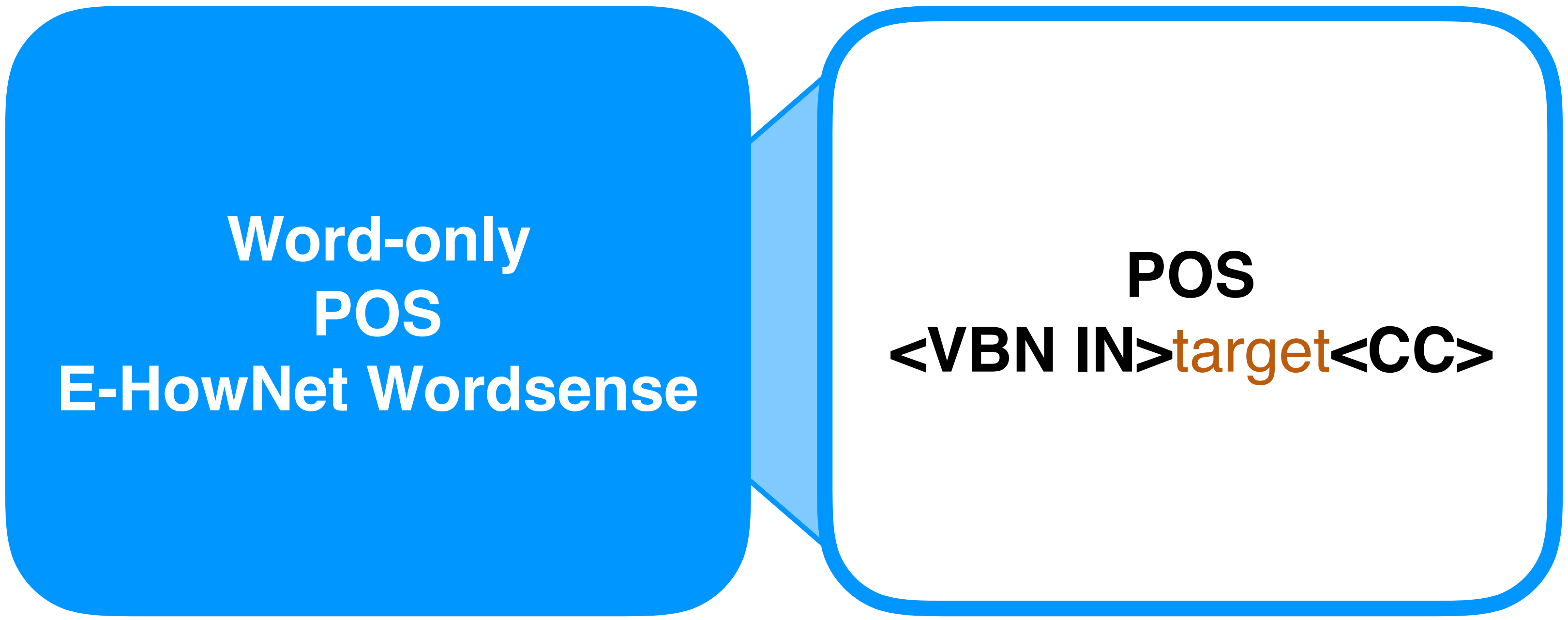
instances

Pattern/Instance Evaluation

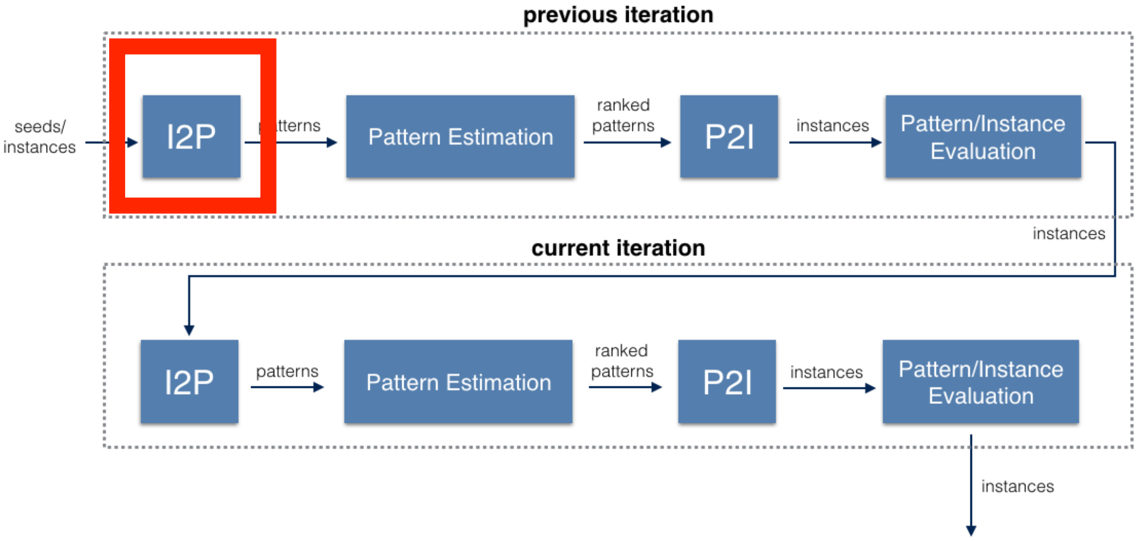
instances

- Pattern Design:

Context


Word-only
POS
E-HowNet Wordsense

POS
<VBN IN>target<CC>



Approach(cont.)

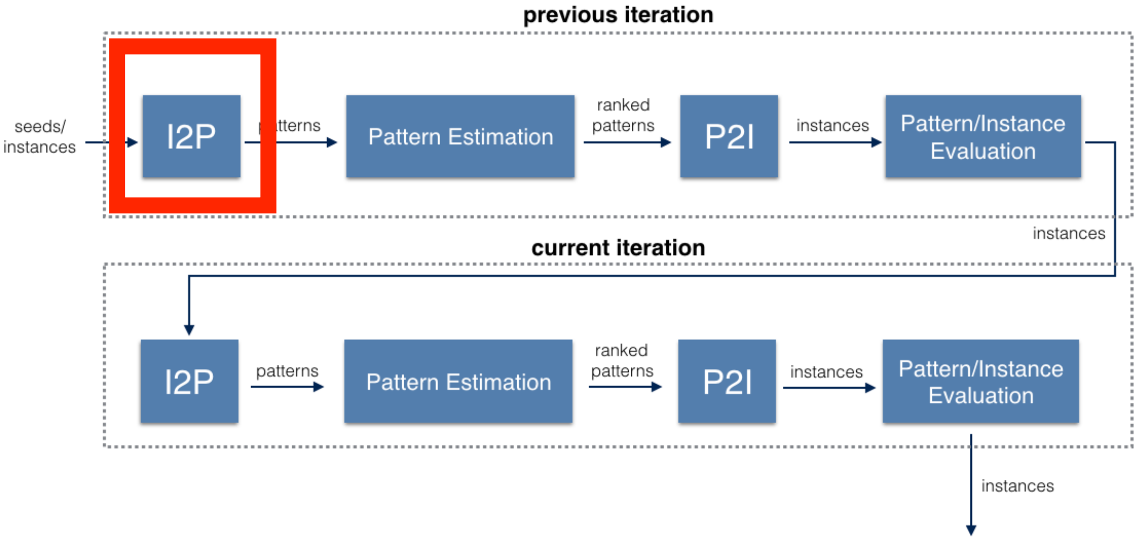
- Pattern Design:

Context

Word-only
POS
E-HowNet Wordsense

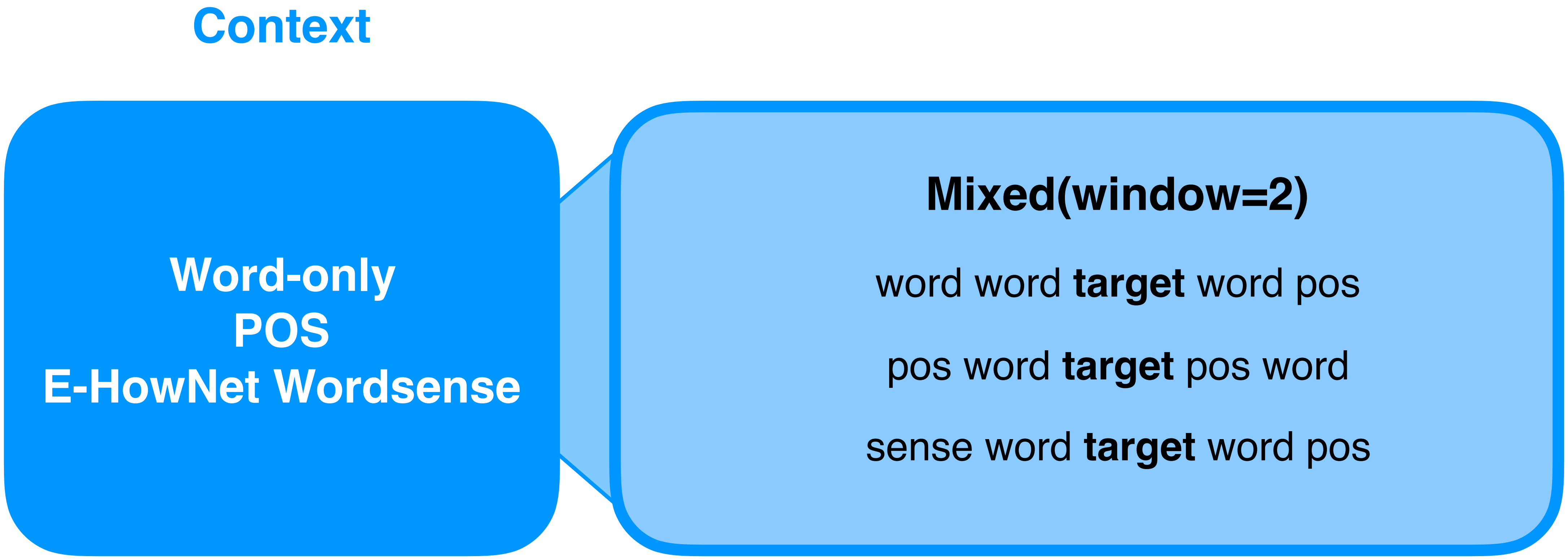
E-HowNet word sense

<human|人.1>target<undertake|擔任.1>



Approach(cont.)

- Pattern Design:



Context

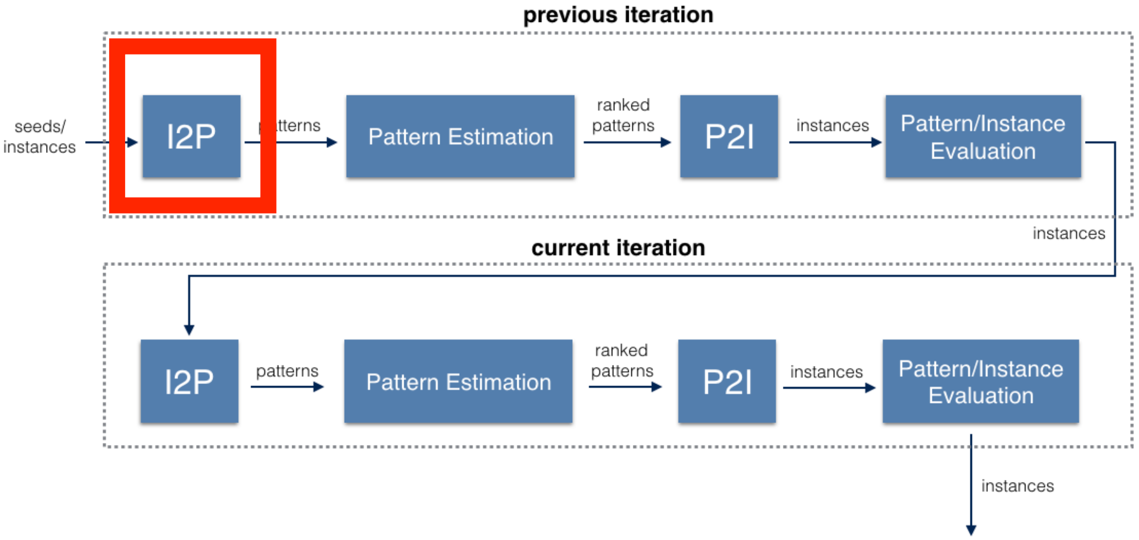
Word-only
POS
E-HowNet Wordsense

Mixed(window=2)

word word **target** word pos

pos word **target** pos word

sense word **target** word pos



Approach(cont.)

- Pattern Design:

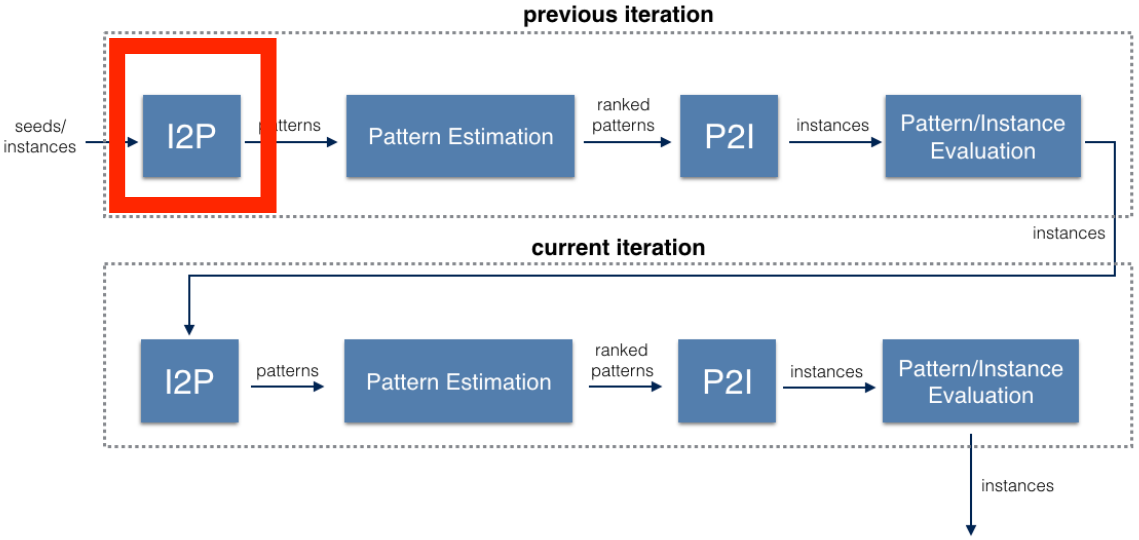
Context

Word-only
POS
E-HowNet Wordsense

Syntactic

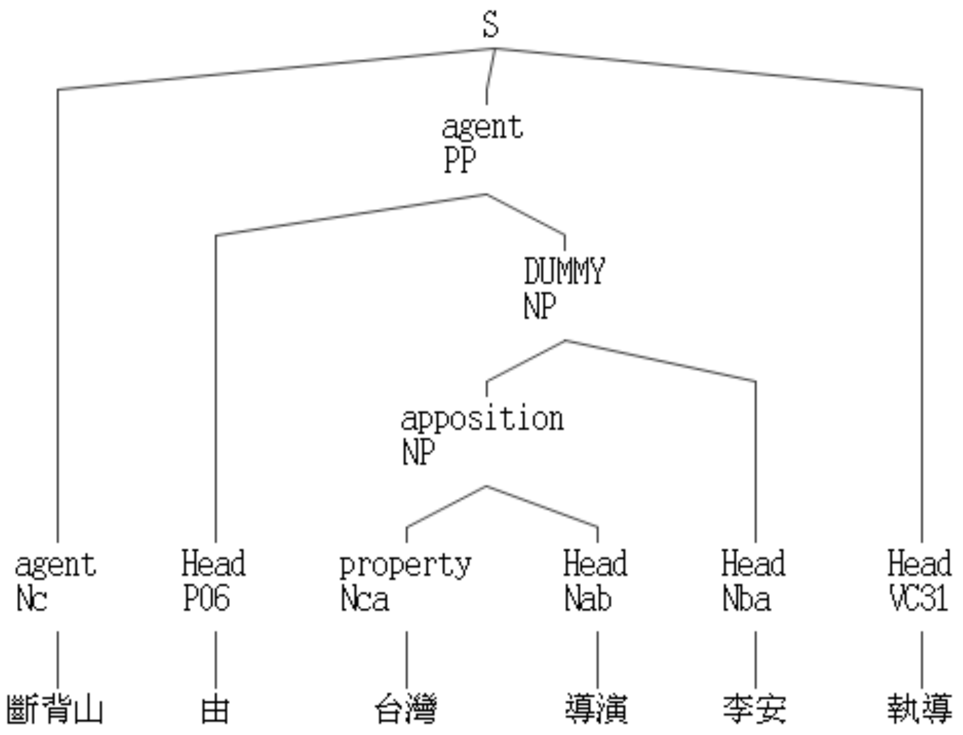
Parse path
Parse path + head

Approach(cont.)



- Pattern Design:

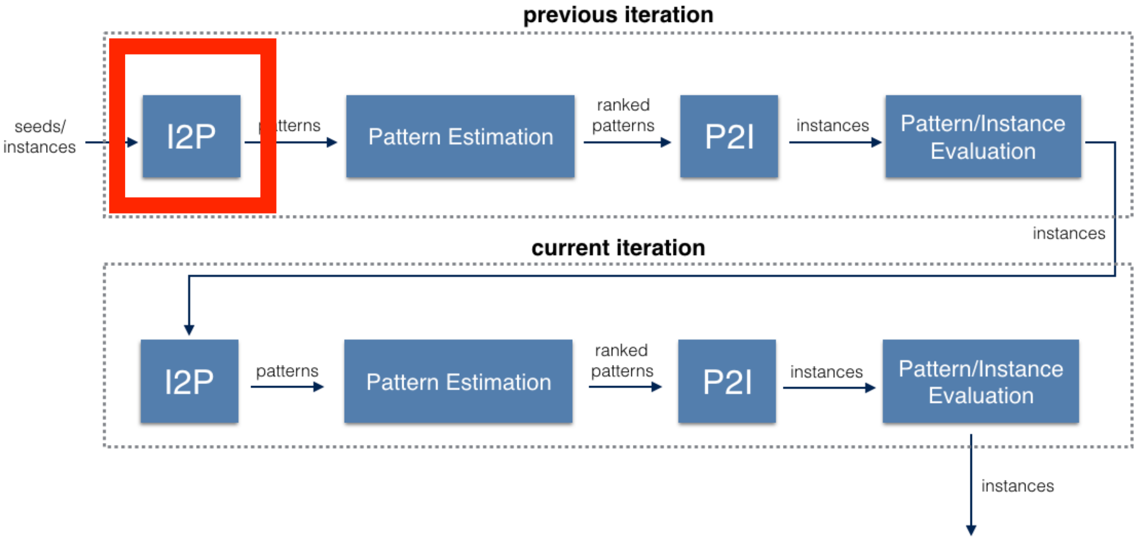
Parser path



root node(S) to seed node(Ang Lee) path：
S -> agent -> DUMMY -> Head

Syntactic

Parse path
Parse path + head


previous iteration

seeds/ instances

I2P

patterns

Pattern Estimation

ranked patterns

P2I

instances

Pattern/Instance Evaluation

instances

current iteration

I2P

patterns

Pattern Estimation

ranked patterns

P2I

instances

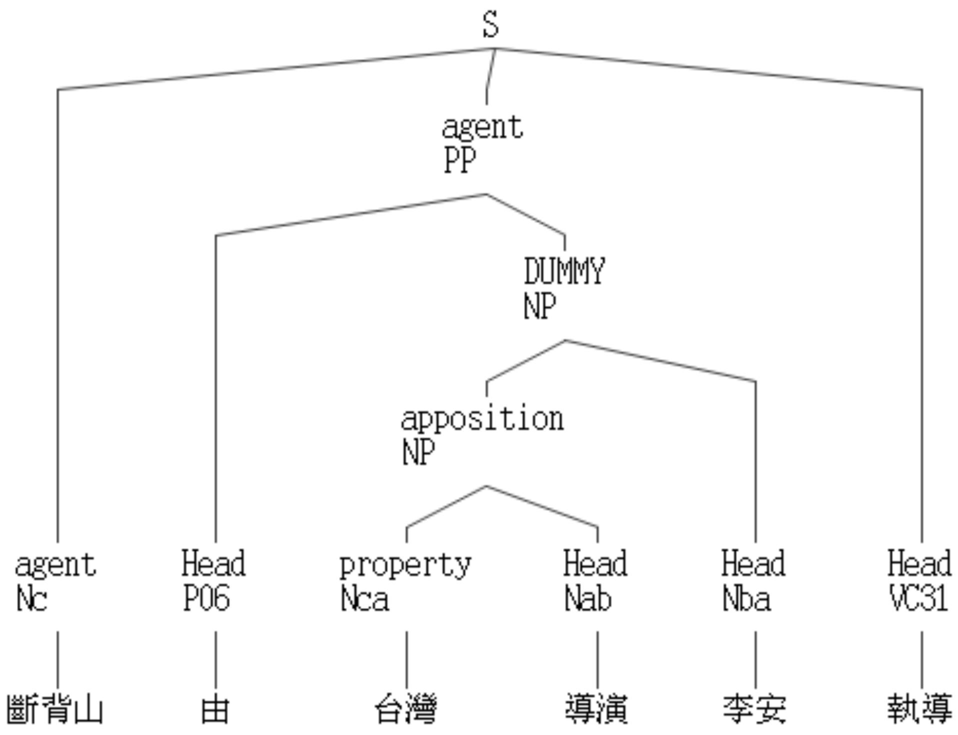
Pattern/Instance Evaluation

instances

Approach(cont.)

- Pattern Design:

Parser path + head



S -> agent -> DUMMY -> Head

(‘directed’, ‘S -> agent -> DUMMY -> Head’)

Syntactic

Parse path

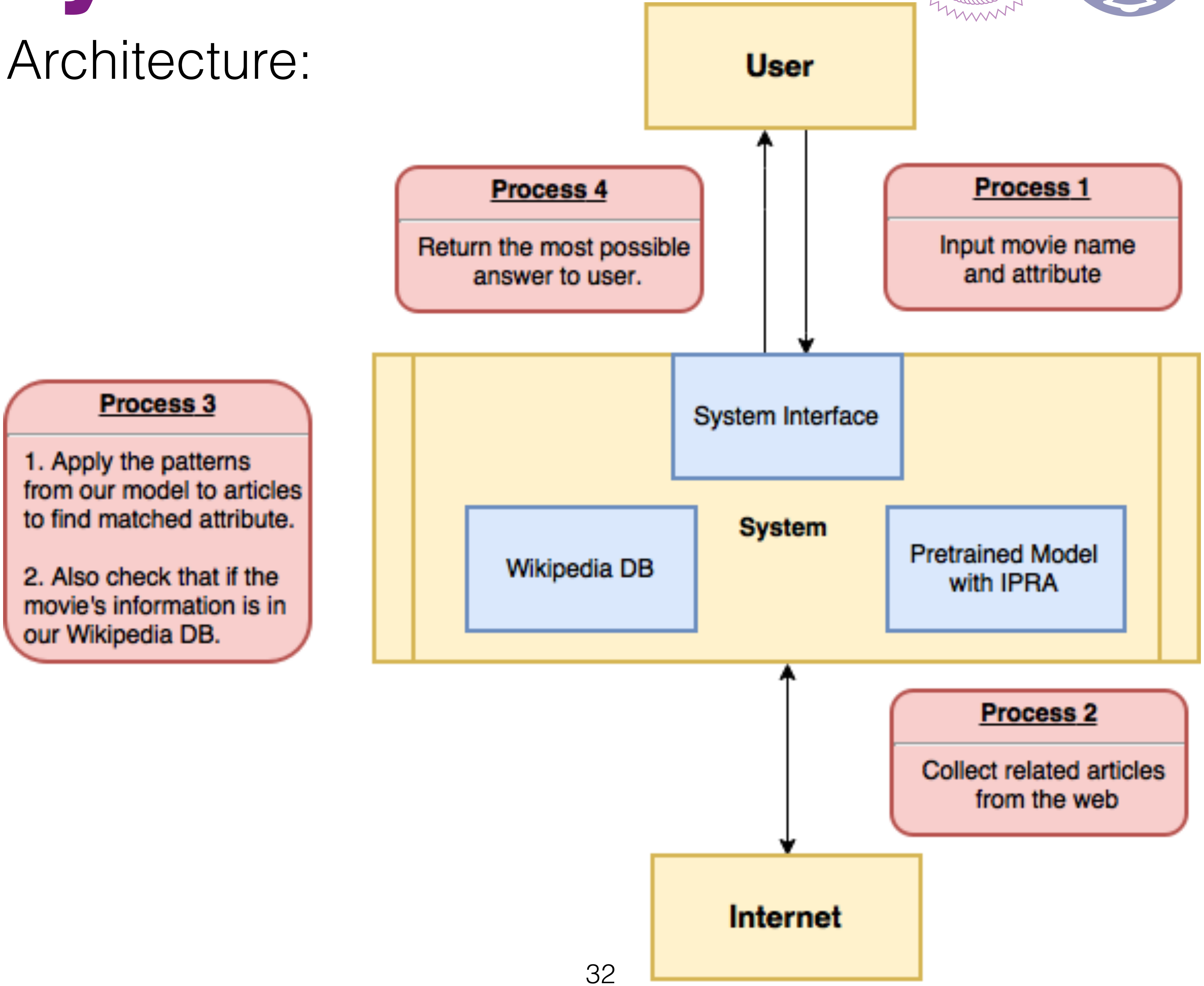
Parse path + head

Outline

- Introduction
- Related Work
- Approach
- System: IExM
- Experiment
- Conclusion

System: IExM

- Architecture:


User
Process 4
Return the most possible answer to user.
Process 1
Input movie name and attribute
Process 3
1. Apply the patterns from our model to articles to find matched attribute.
2. Also check that if the movie's information is in our Wikipedia DB.
System Interface
System
Wikipedia DB
Pretrained Model with IPRA
Process 2
Collect related articles from the web
Internet

System: IExM(cont.)

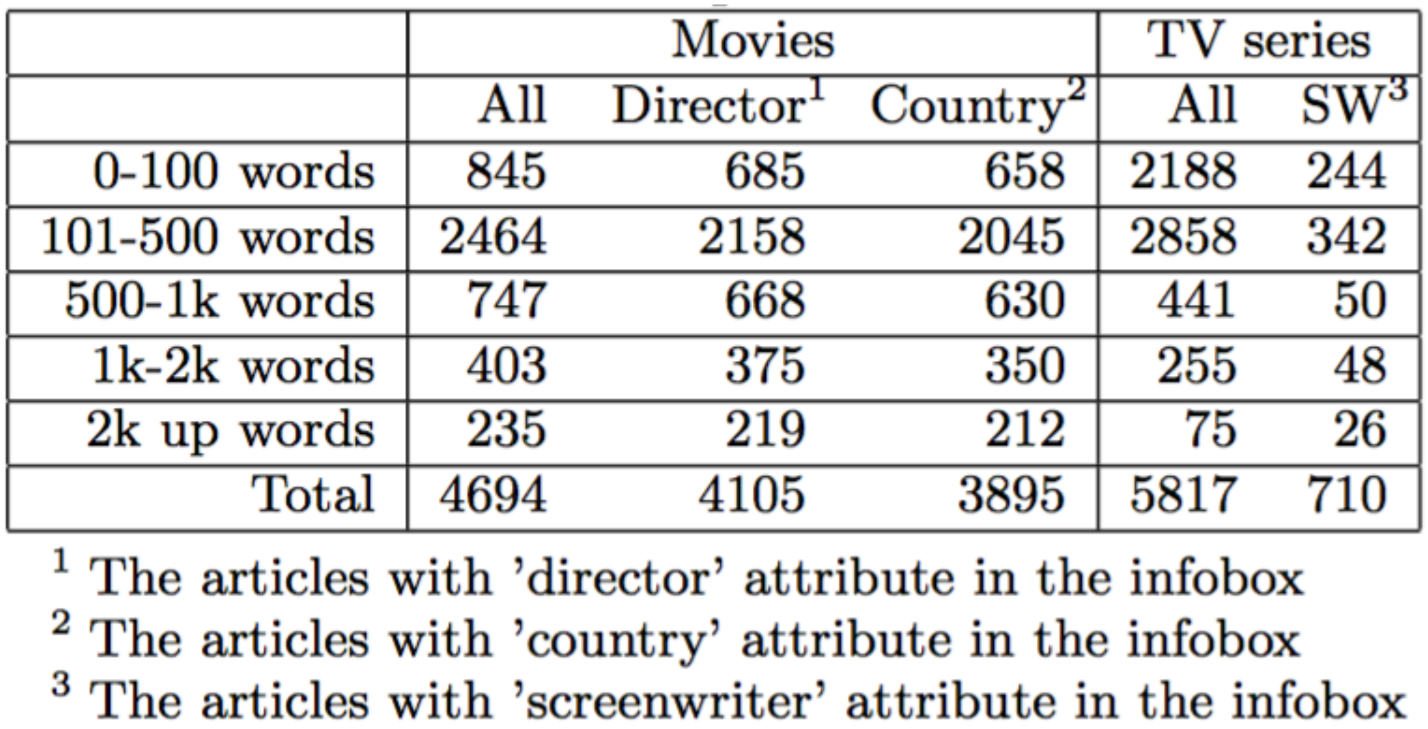
- Demo:
  - IExM: http://learn.iis.sinica.edu.tw/IExM

Outline

- Introduction
- Related Work
- Approach
- System: IExM
- **Experiment**
- Conclusion

Experiment

- Data Set: Wikipedia



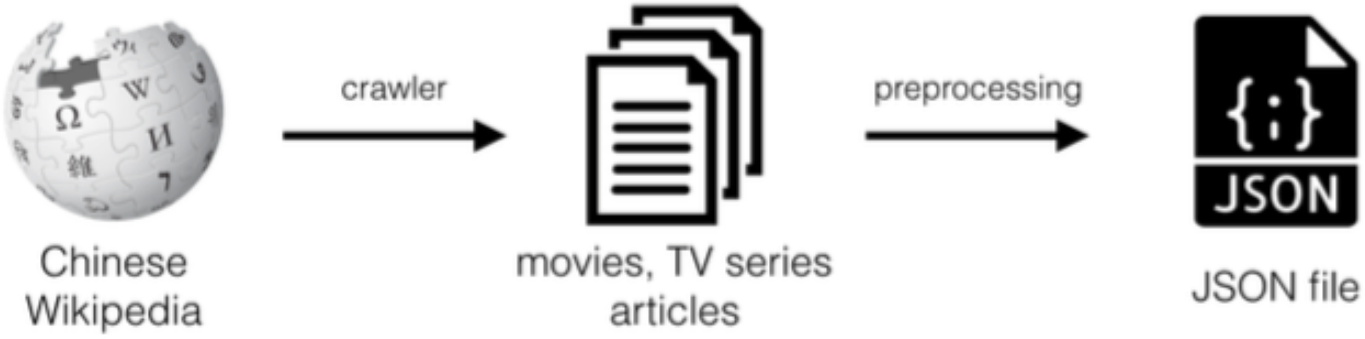
	Movies			TV series	
	All	Director[1]	Country[2]	All	SW[3]
0-100 words	845	685	658	2188	244
101-500 words	2464	2158	2045	2858	342
500-1k words	747	668	630	441	50
1k-2k words	403	375	350	255	48
2k up words	235	219	212	75	26
Total	4694	4105	3895	5817	710

[1] The articles with 'director' attribute in the infobox

[2] The articles with 'country' attribute in the infobox

[3] The articles with 'screenwriter' attribute in the infobox

- Data preprocessing flow



Experiment(cont.)

- Compare Pattern types:

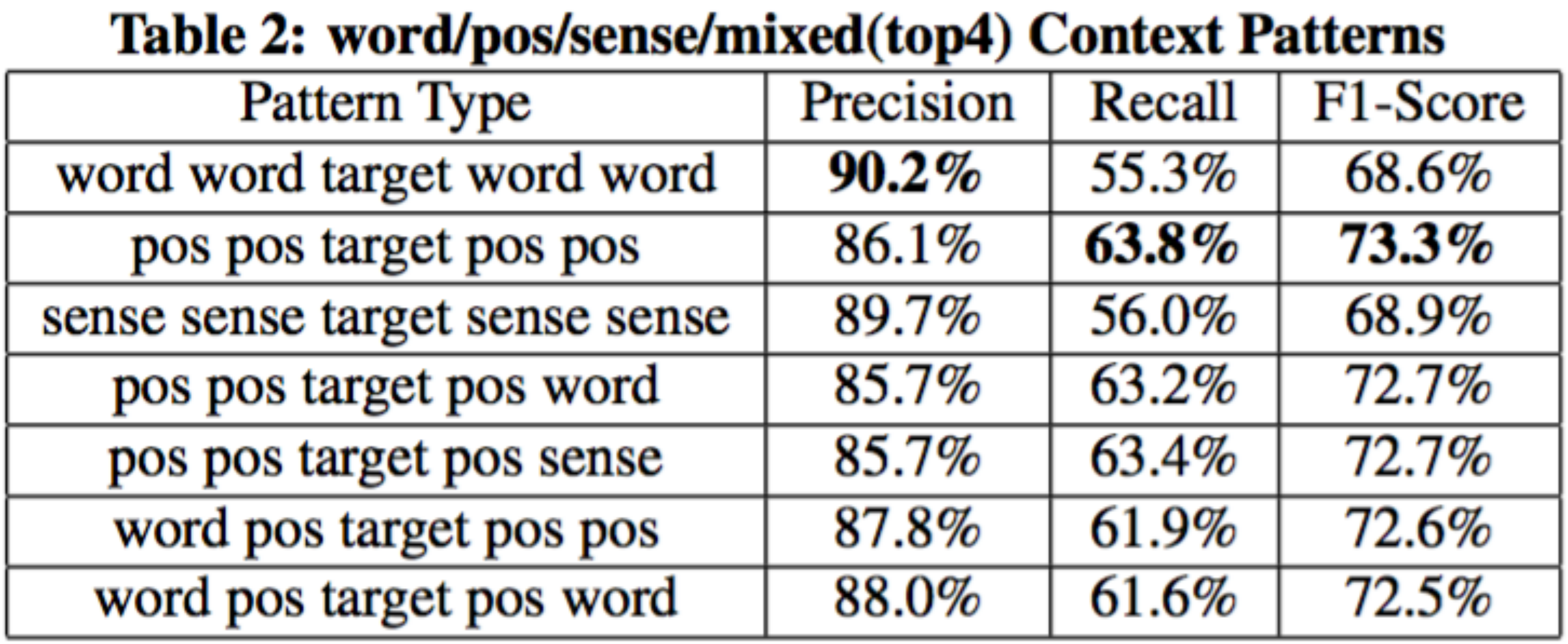


Table 2: word/pos/sense/mixed(top4) Context Patterns

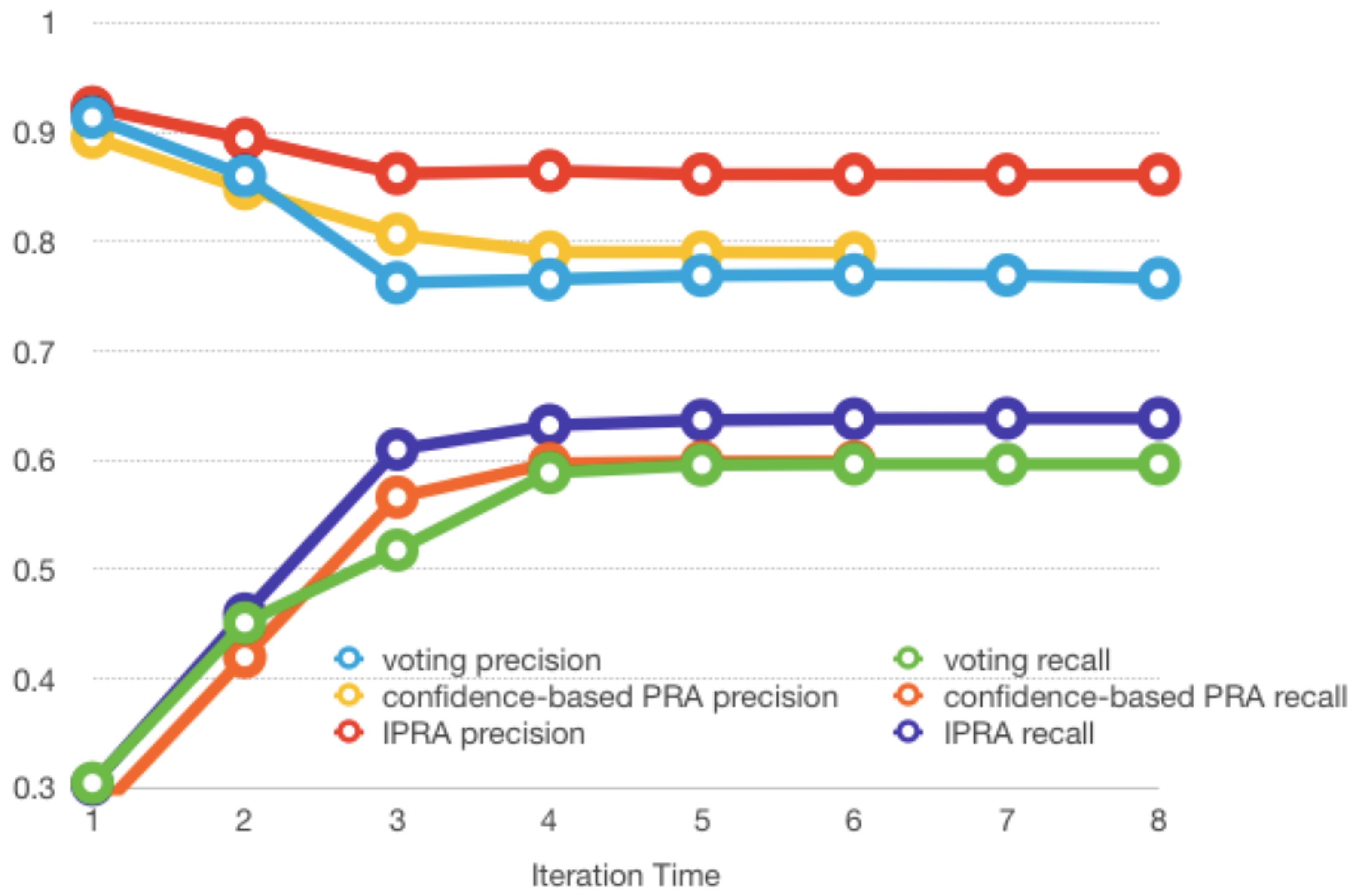
Pattern Type	Precision	Recall	F1-Score
word word target word word	**90.2%**	55.3%	68.6%
pos pos target pos pos	86.1%	**63.8%**	**73.3%**
sense sense target sense sense	89.7%	56.0%	68.9%
pos pos target pos word	85.7%	63.2%	72.7%
pos pos target pos sense	85.7%	63.4%	72.7%
word pos target pos pos	87.8%	61.9%	72.6%
word pos target pos word	88.0%	61.6%	72.5%

Experiment(cont.)

- Choose best pattern type: pos pos target pos pos
- Compare Algorithms:
  - Voting: collecting the instances extracted by new patterns in one article, and then decide the instance with highest votes
  - Confidence-Based Pattern Ranking Algorithm (PRA): considering only the confidence of pattern to estimate the patterns' quality, that is, using only equation of precision and equation of confidence.
  - IPRA: our work

Experiment(cont.)

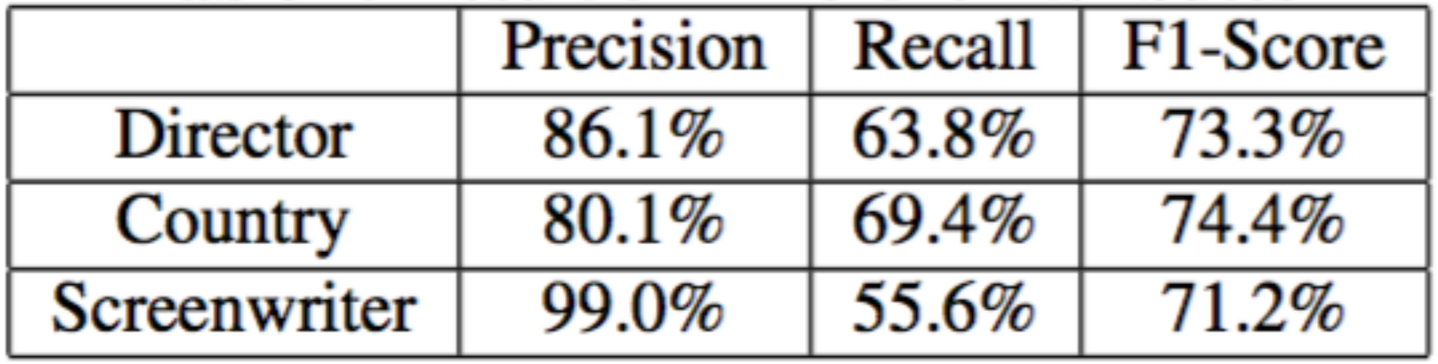
- Choose best pattern type: pos pos target pos pos
- Compare Algorithms:
    - Voting: (f1-score:0.670)
    - Confidence-Based PRA: (f1-score: 0.680)
    - IPRA: our work (f1-score: 0.733)



Experiment(cont.)

- Compare Different Attributes:

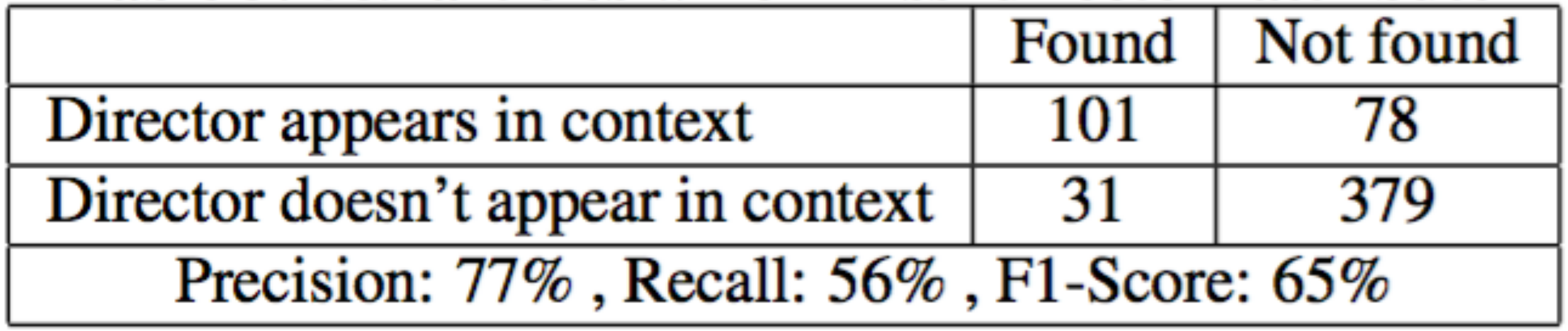
Table 1: Result of Different Attributes



	Precision	Recall	F1-Score
Director	86.1%	63.8%	73.3%
Country	80.1%	69.4%	74.4%
Screenwriter	99.0%	55.6%	71.2%

- Articles which miss ‘director’ attribute

Table 3: 589 articles which miss ’director’ attribute



	Found	Not found
Director appears in context	101	78
Director doesn’t appear in context	31	379
Precision: 77% , Recall: 56% , F1-Score: 65%		

Outline

- Introduction
- Related Work
- Approach
- System: IExM
- Experiment
- **Conclusion**

Conclusion

- For improving pattern ranking
  - We propose a new distant-supervised learning framework which is able to dynamically estimate and rank all generated patterns based on their application.
- As more patterns are generated and ranked, the coverage and precision of extracted instances can be gradually improved and then achieve a high performance in the end.
- Future work:
  - Integrate coupled training with a large amount of couples relations