

Mayank Kejriwal

University of Southern California



Lexicons and reference sets - I

Outline

- ☒ Introduction
- ☐ Alignment
- ☐ Extraction
- ☐ Results
- ☐ Discussion

Ungrammatical & Unstructured Text

Page 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

	Topic	Replies	Last Comment	Started By
	 SACRAMENTO HOTEL LIST	0	11/21/04 9:56 pm	westcoastman
	3* Rancho Cordova Holiday Inn \$35, 1 nite (12/11)	1	12/9/04 12:37 am	future canadian
	3* Doubletree Sacto Arden 12/11 1 Night \$34	1	12/7/04 4:46 pm	OCTraveler
	4* Sacramento Failed Bid \$85 12/7	1	12/6/04 6:29 pm	Sheryl
	Failed bid Sacramento Downtown 12/6 for 1 night, 4*	13	12/6/04 6:25 pm	emaij
	2.5* Wingate Inn Rancho Cordova 5/10-5/13/05 \$32	0	12/4/04 7:11 pm	ego68
	3* DoubleTree Sacramento \$35 (12/04/04)	0	11/30/04 11:34 pm	shizzolator
	2.5* Rancho Cordova Wingate Inn \$32 (11/23-25)	1	11/27/04 12:19 pm	Profiler
	4* DT Hyatt 11/21 \$60 11/23 \$60; Sheraton Grand 11/25 \$55	0	11/22/04 1:22 pm	bonish
	3* Doubletree Arden/Sacramento \$37 11/19	1	11/20/04 1:53 am	ahallez
	2.5* Wingate Inn Rancho Cordova \$33 11/13	2	11/19/04 1:44 am	cykick42
	2.5* DT Hawthorne Suites \$40 (11/18-20)	0	11/18/04 10:08 pm	Colfax30
	Roseville 2.5*Larkspur \$72(11/22-24) 2* Fairfield \$80(11/24)	2	11/17/04 4:38 pm	mcrinca
	3* Rancho Cordova Holiday Inn \$32 (11/17)	0	11/16/04 10:20 pm	Colfax30
	3* Doubletree Sacramento \$40 (11/11)	2	11/16/04 11:05 am	OCTraveler
	3* Doubletree Sacramento Arden \$36 11/24	0	11/15/04 1:04 am	bomawin

Ungrammatical & Unstructured Text

For simplicity → “posts”

Goal:

<hotelArea>univ. ctr.</hotelArea>		
Beware 2* at the airport!!!!	2	7/18/00 1:25 am
\$25 winning bid at holiday inn sel univ. ctr.	1	6/26/00 1:48 pm
3* Holiday Inn North-McKnight Rd, \$10+20, 1/19	3	1/27/01 6:34 pm

<price>\$25</price> <hotelName>holiday inn sel.</hotelName>

Wrapper based IE does not apply (e.g. Stalker, RoadRunner)

NLP based IE does not apply (e.g. Rapier)


Reference Sets


IE infused with outside knowledge

“Reference Sets”

- Collections of known entities and the associated attributes
- Online (offline) set of docs
 - CIA World Fact Book
- Online (offline) database
 - Comics Price Guide, Edmunds, etc.
- Build from ontologies on Semantic Web

Comics Price Guide Reference Set



**CGC**
Comics Guaranty, LLC

Submit your books online
and get **20% off**

[CONTACT US](#)
[MEDIA KIT](#)
[ADMIN LOGIN](#)
[AD MANAGE](#)

[HOME](#) [GRADING](#) [MESSAGE BOARDS](#) [STORE](#) [CLASSIFIEDS](#) [AUCTIONS](#) [ISSUES SALES](#) [FAQ](#)

Login 131 users

Username:

Password:

☐ Remember Me [Forgot Login](#) [Sign Up](#)

[Login](#)


SEARCH BY PUBLISHER


SEARCH BY KEYWORDS


[Search](#) [Search](#)


[Marvel](#) [# A B C D E F G H I J K L M N O P Q R S T U V W X Y Z](#)

FANTASTIC FOUR (1961-1996,2003-CURRENT) 255349 Total Searches

 **Add To Collection**
books you do have

 **Add To Want List**
books you must have


 **View Collection**
see the issues you own


 **Print This**
take home copy

Select All ☐

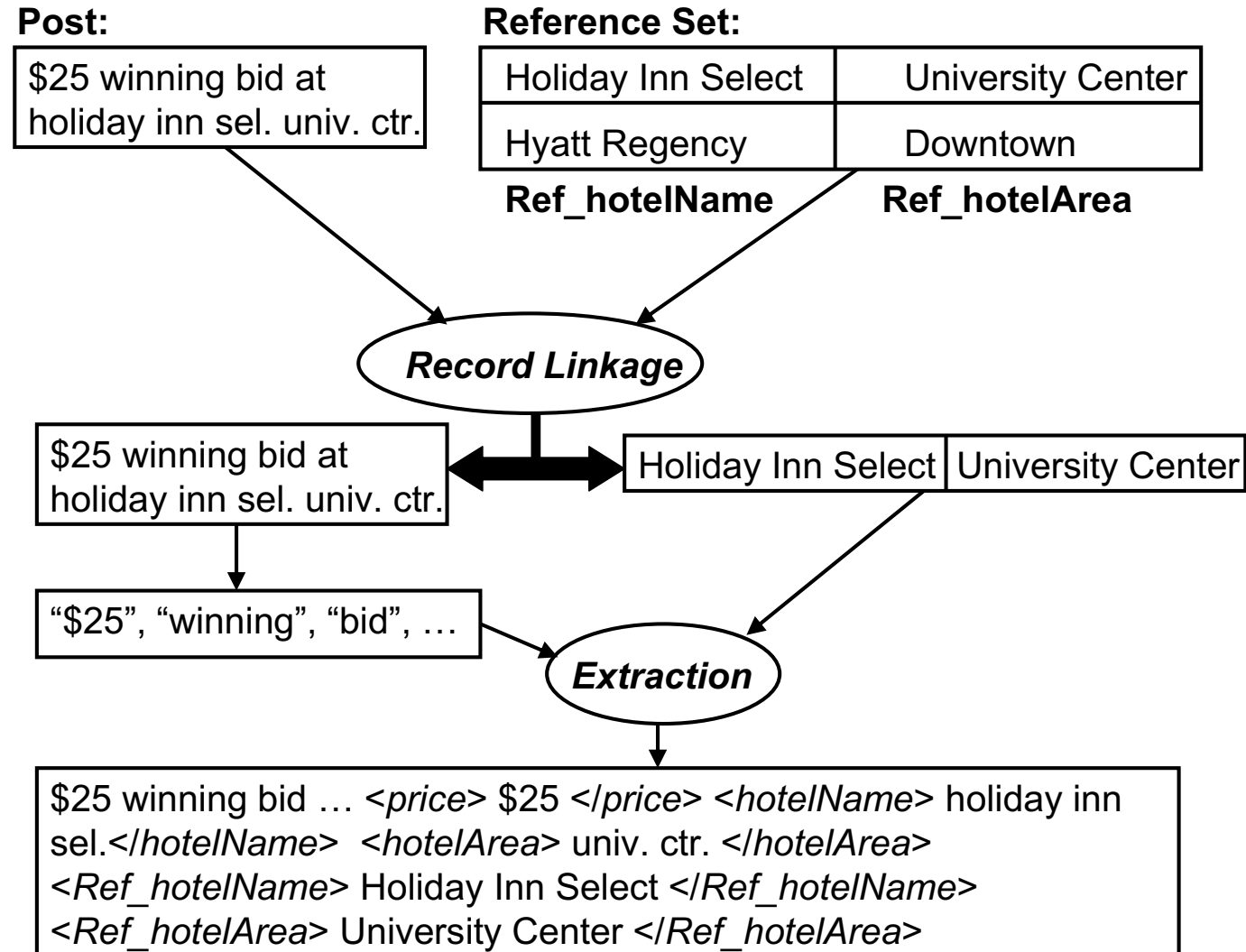
Page [1](#) [2](#) [3](#) [4](#) [5](#) [6](#)

[Find Issue](#)

Issue #	9.4 Value	9.4 CGC Graded	For Sale	Cover
<input type="checkbox"/> # 1	 \$32,000.00	\$192,000.00		VIEW
First Appearance: Fantastic Four and The Mole Man				
<input type="checkbox"/> # 1A	\$300.00	\$1,800.00	SALE	VIEW
Golden Record Reprint Edition				
<input type="checkbox"/> # 1B	\$200.00	\$1,200.00		VIEW
Comic removed from album				
<input type="checkbox"/> # 2	\$5,250.00	\$31,500.00		VIEW
First Appearance: The Skrulls				
<input type="checkbox"/> # 3	\$3,000.00	\$18,000.00		VIEW
First Fantastic Four Costume				



Algorithm Overview – Use of Ref Sets

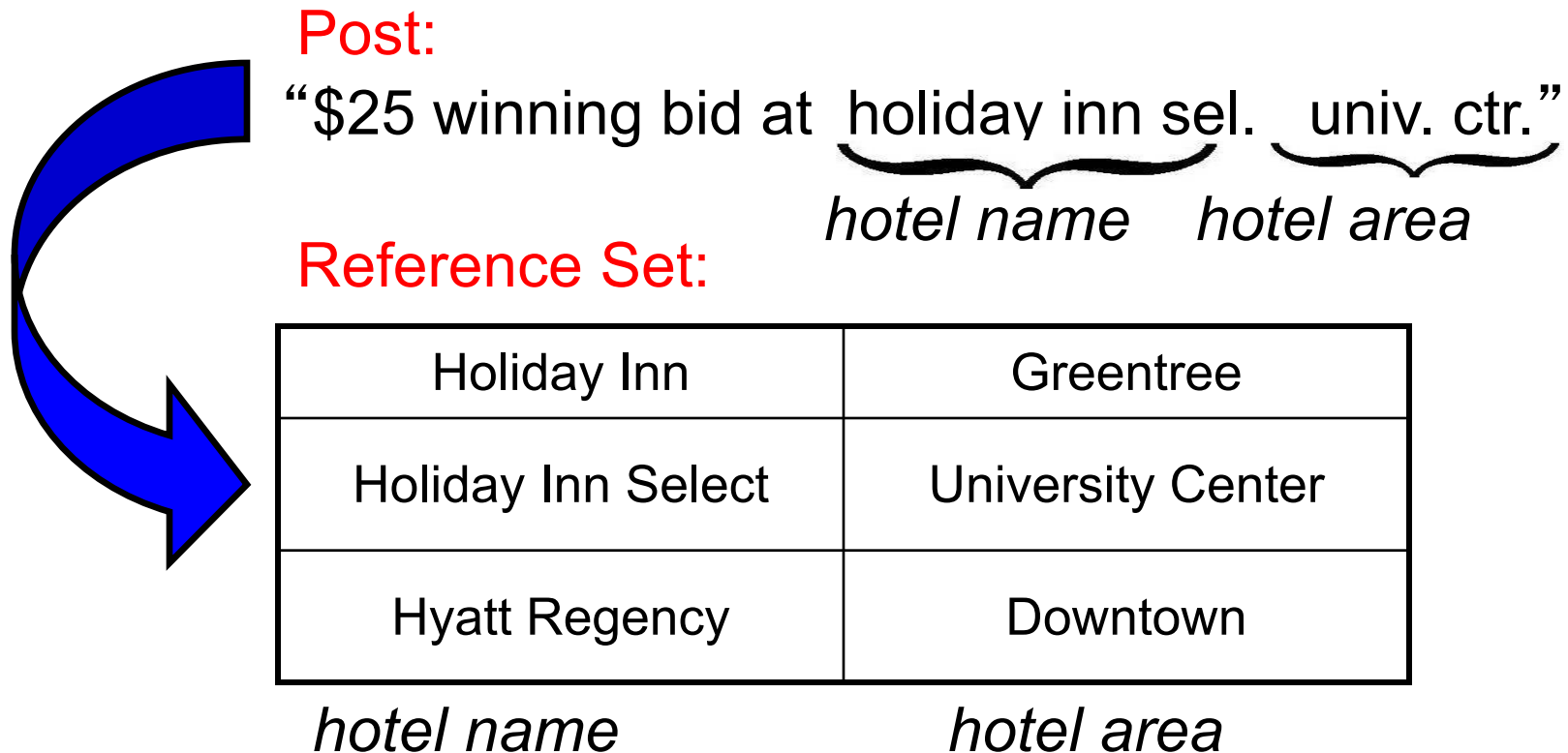


Outline

- ☐ Introduction
- ☐ Alignment
- ☐ Extraction
- ☐ Results
- ☐ Discussion

Our Record Linkage Problem

- *Posts not yet decomposed attributes.*
- *Extra tokens that match nothing in Ref Set.*



Our Record Linkage Solution

P = "\$25 winning bid at holiday inn sel. univ. ctr."

Record Level Similarity + Field Level Similarities

$V_{RL} = \langle RL_scores(P, \text{"Hyatt Regency Downtown"}),$
 $RL_scores(P, \text{"Hyatt Regency"}),$
 $RL_scores(P, \text{"Downtown"}) \rangle$

Binary Rescoring

SVM

Best matching member of the reference set for the post

Last Alignment Step

Return reference set attributes as annotation for the post

Post:

\$25 winning bid at holiday inn sel. univ. ctr.

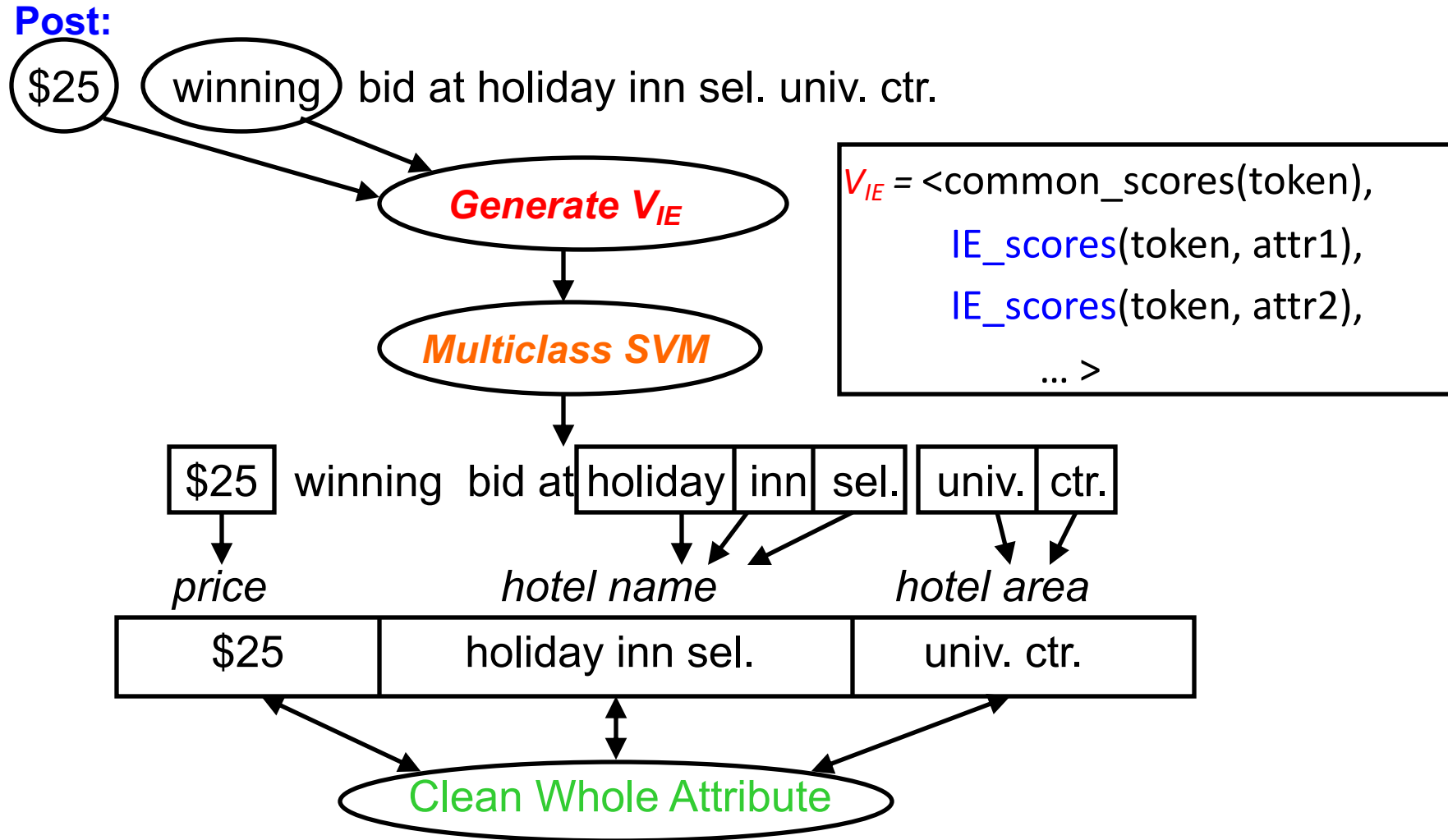
<Ref_hotelName>Holiday Inn Select</Ref_hotelName>

<Ref_hotelArea>University Center</Ref_hotelArea>

Outline

- ☐ Introduction
- ☐ Alignment
- ☐ Extraction
- ☐ Results
- ☐ Discussion

Extraction Algorithm



Common Scores

- Some attributes not in reference set
 - Reliable characteristics
 - Infeasible to represent in reference set
 - E.g. prices, dates
- Can use characteristics to extract/annotate these attributes
 - Regular expressions, for example
- These types of scores are what compose ***common_scores***