



Semantic annotation of unstructured and ungrammatical text

Matthew Michelson and Craig A. Knoblock
Information Sciences Institute
Department of Computer Science
University of Southern California

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	westcoastma
3* Rancho Cordova Holiday Inn \$35, 1 nite (12/11)	1	12/9/04 12:37 am	future canad
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





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

CGC, The Official Grading Service of CPG

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](#)

SEARCH BY PUBLISHER

SEARCH BY KEYWORDS

[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

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				

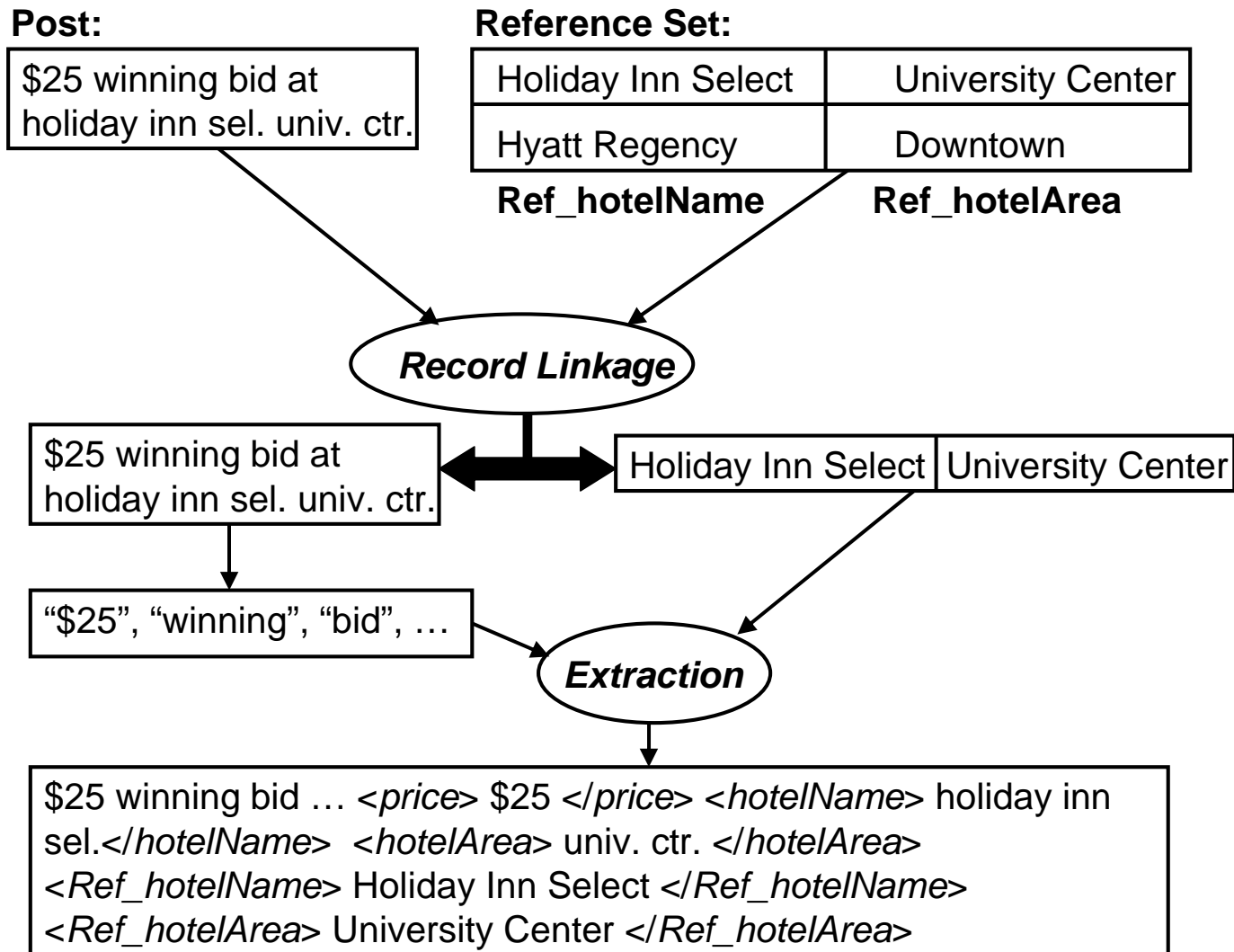




2 Step Approach to Annotation

1. Align post to a member of the reference set
2. Exploit the matching member of reference set for extraction/annotation

Algorithm Overview – Use of Ref Sets



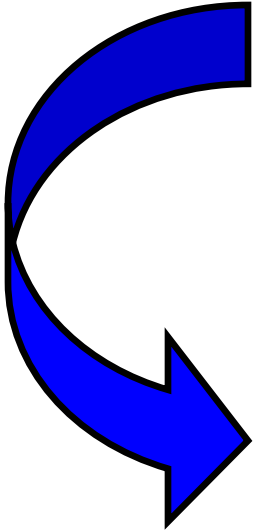
Our Record Linkage Problem

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

Post:

“\$25 winning bid at holiday inn sel. univ. ctr.”
hotel name hotel area

Reference Set:



Holiday Inn	Greentree
Holiday Inn Select	University Center
Hyatt Regency	Downtown

hotel name

hotel area

Our Record Linkage Solution

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

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

RL_scores

RL_scores(s, t)

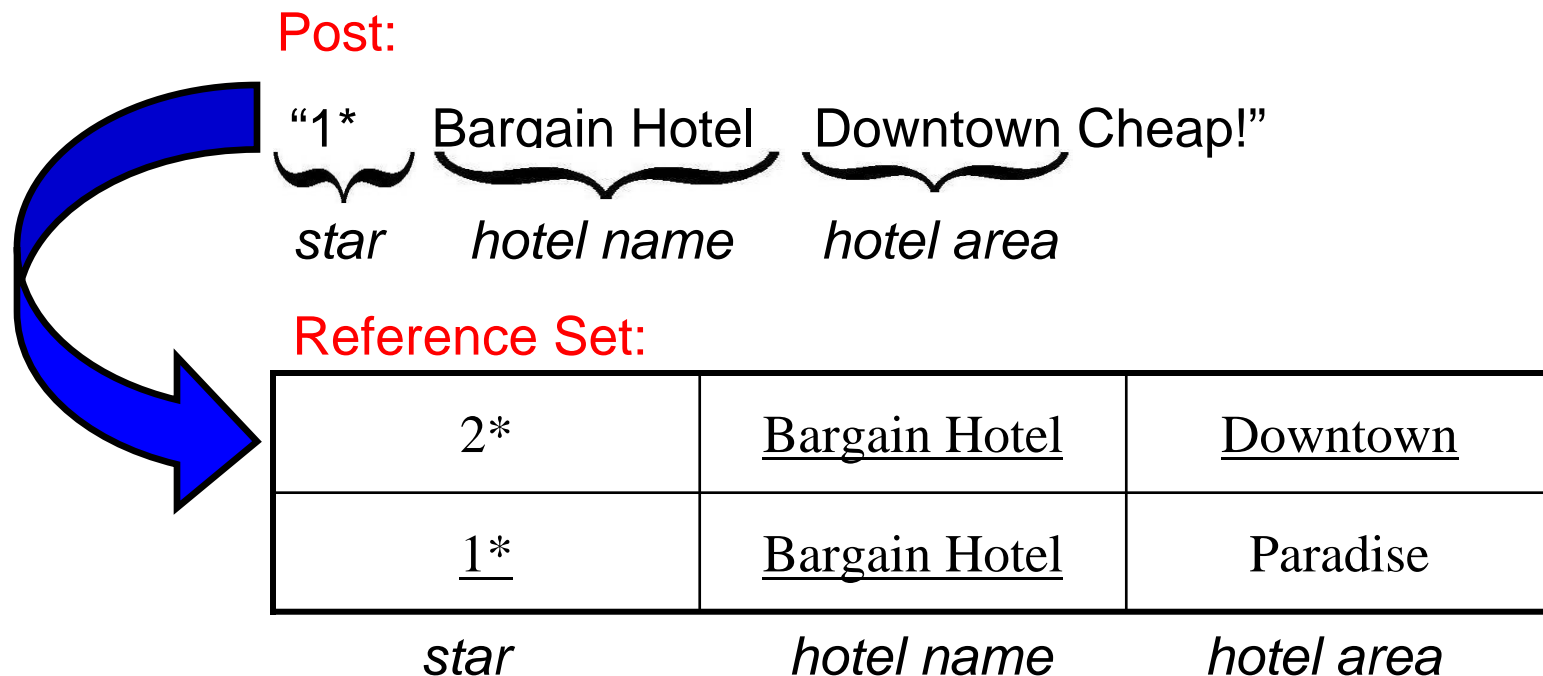
< token_scores(s, t), edit_scores(s, t), other_scores(s, t) >

Jensen-Shannon
(Dirichlet & Jelenik-Mercer)
Jaccard

Levenstein
Smith-Waterman
Jaro-Winkler

Soundex
Porter Stemmer

Record Level Similarity Problem



What if equal *RLS* but different attributes? Many more hotels share **Star** than share **Hotel Area** → need to reflect **Hotel Area** similarity more discriminative...

Binary Rescoring

$$\textit{Candidates} = \langle V_{RL1}, V_{RL2}, \dots, V_{RLn} \rangle$$

$V_{RL}(s)$ with max value at index i set that value to 1. All others set to 0.

$$V_{RL1} = \langle 0.999, 1.2, \dots, 0.45, 0.22 \rangle$$

$$V_{RL2} = \langle 0.888, 0.0, \dots, 0.65, 0.22 \rangle$$



$$V_{RL1} = \langle 1, 1, \dots, 0, 1 \rangle$$

$$V_{RL2} = \langle 0, 0, \dots, 1, 1 \rangle$$

Emphasize best match →
similarly close values but only one is best match



SVM Classification

Support Vector Machine (SVM)

- Trained to classify matches/ non-matches
- Returns score from decision function
- Best Match: Candidate that is a match & max. score from decision function
 - 1-1 mapping: If more than one cand. with max. score → throw them all away
 - 1-N mapping: If more than one cand. with max. score → keep first one or keep random one w/in set of max.



Last Alignment Step

Return reference set attributes as annotation for the post

Post:

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

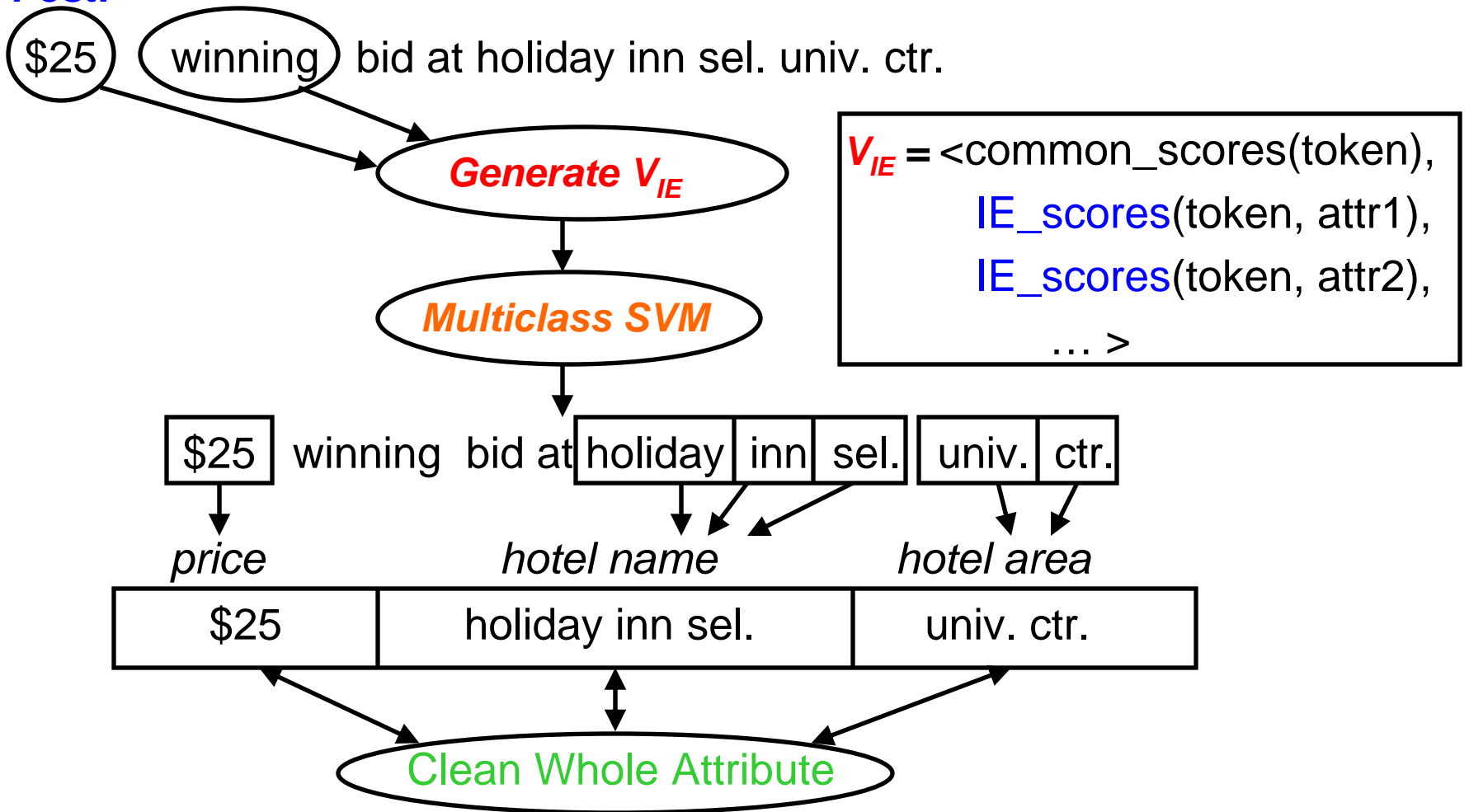
<Ref_hotelName>Holiday Inn Select</Ref_hotelName>

<Ref_hotelArea>University Center</Ref_hotelArea>

... discuss implications a little later...

Extraction Algorithm

Post:





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*

Cleaning an attribute: Example

Baseline scores: *holiday inn sel. in*

Jaro-Winkler (edit): 0.87

Jaccard (token): 0.4

Iteration 1

Scores: ~~*holiday inn sel. in*~~

Jaro-Winkler (edit): 0.92 (> 0.87) Jaccard (token): 0.5 (> 0.4)

New Hotel Name: *holiday inn sel.*

New baselines

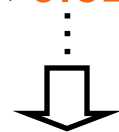
Iteration 2

Scores: ~~*holiday inn sel.*~~

Jaro-Winkler (edit): 0.84 (< 0.92) Jaccard (token): 0.66 (> 0.5)

Scores: ~~*holiday inn sel.*~~

Jaro-Winkler (edit): 0.87 (< 0.92) Jaccard (token): 0.25 (< 0.5)



No improvement → terminate

holiday inn sel.



Experimental Data Sets

Hotels

□ *Posts*

- 1125 posts from www.biddingfortravel.com
 - Pittsburgh, Sacramento, San Diego
 - Star rating, hotel area, hotel name, price, date booked

□ *Reference Set*

- 132 records
- Special posts on BFT site.
 - Per area – list any hotels ever bid on in that area
 - Star rating, hotel area, hotel name



Experimental Data Sets

Comics

□ *Posts*

- 776 posts from EBay
 - “Incredible Hulk” and “Fantastic Four” in comics
 - Title, issue number, price, condition, publisher, publication year, description (1st appearance the Rhino)

□ *Reference Sets*

- 918 comics, 49 condition ratings
- Both come from ComicsPriceGuide.com
 - For FF and IH
 - Title, issue number, description, publisher



Comparison to Existing Systems

Our Implementation

- Phoebus

Record Linkage

- WHIRL
 - RL allows non-decomposed attributes

Information Extraction

- Simple Tagger (CRF)
 - State-of-the-art IE
- Amilcare
 - NLP based IE

Record linkage results

	Prec.	Recall	F-Measure
Hotel			
Phoebus	93.60	91.79	92.68
WHIRL	83.52	83.61	83.13
Comic			
Phoebus	93.24	84.48	88.64
WHIRL	73.89	81.63	77.57

10 trials – 30% train, 70% test

Token level Extraction results: Hotel domain

		Prec.	Recall	F-Measure	Freq
<i>Area</i>	Phoebus	89.25	87.50	88.28	809.7
	Simple Tagger	92.28	81.24	86.39	
	Amilcare	74.2	78.16	76.04	
<i>Date</i>	Phoebus	87.45	90.62	88.99	751.9
	Simple Tagger	70.23	81.58	75.47	
	Amilcare	93.27	81.74	86.94	
<i>Name</i>	Phoebus	94.23	91.85	93.02	1873.9
	Simple Tagger	93.28	93.82	93.54	
	Amilcare	83.61	90.49	86.90	
<i>Price</i>	Phoebus	98.68	92.58	95.53	850.1
	Simple Tagger	75.93	85.93	80.61	
	Amilcare	89.66	82.68	85.86	
<i>Star</i>	Phoebus	97.94	96.61	97.84	766.4
	Simple Tagger	97.16	97.52	97.34	
	Amilcare	96.50	92.26	94.27	

Not Significant

Token level Extraction results: Comic domain

		Prec.	Recall	F-Measure	Freq
<i>Condition</i>	Phoebus	91.8	84.56	88.01	410.3
	Simple Tagger	78.11	77.76	77.80	
	Amilcare	79.18	67.74	72.80	
<i>Descript.</i>	Phoebus	69.21	51.50	59.00	504.0
	Simple Tagger	62.25	79.85	69.86	
	Amilcare	55.14	58.46	56.39	
<i>Issue</i>	Phoebus	93.73	86.18	89.79	669.9
	Simple Tagger	86.97	85.99	86.43	
	Amilcare	88.58	77.68	82.67	
Price	Phoebus	80.00	60.27	68.46	10.7
	Simple Tagger	84.44	44.24	55.77	
	Amilcare	60.00	34.75	43.54	

Token level Extraction results: Comic domain (cont.)

		Prec.	Recall	F-Measure	Freq
<i>Publisher</i>	Phoebus	83.81	95.08	89.07	61.1
	Simple Tagger	88.54	78.31	82.83	
	Amilcare	90.82	70.48	79.73	
<i>Title</i>	Phoebus	97.06	89.90	93.34	1191.1
	Simple Tagger	97.54	96.63	97.07	
	Amilcare	96.32	93.77	94.98	
Year	Phoebus	98.81	77.60	84.92	120.9
	Simple Tagger	87.07	51.05	64.24	
	Amilcare	86.82	72.47	78.79	

Summary extraction results

Expensive to label training data...

	Prec.	Recall	F-Mes.	# Train.
Hotel (30%)	93.6	91.79	92.68	338
Hotel (10%)	93.66	90.93	92.27	113
Comic (30%)	93.24	84.48	88.64	233
Comic (10%)	91.41	83.63	87.34	78

Token Level

Hotel (30%)	87.44	85.59	86.51
Hotel (10%)	86.52	84.54	85.52
Comic (30%)	81.73	80.84	81.28
Comic (10%)	79.94	76.71	78.29

Field Level



Reference Set Attributes as Annotation

- Standard query values
- Include info not in post
 - If post leaves out “Star Rating” can still be returned in query on “Star Rating” using reference set annotation
- Perform better at annotation than extraction
 - Consider record linkage results as field level extraction
 - E.g., no system did well extracting comic desc.
 - +20% precision, +10% recall using record link



Reference Set Attributes as Annotation

Then why do extraction at all?

- Want to see actual values
- Extraction can annotate when record linkage is wrong
 - Better in some cases at annotation than record linkage
 - If wrong record matched, usually close enough record to get some extraction parts right
- Learn what something is not
 - Helps to classify things not in reference set
 - Learn which tokens to ignore better



Related Work

- Generate mark-up for Semantic Web
 - Rely on lexical info (e.g. S-CREAM, MnM) or structure (ADEL)
- Record Linkage
 - Require decomposed attributes
 - WHIRL is exception, used in experiments
- Data Cleaning
 - Tuple-to-tuple transformations (Fuzzy Match Similarity)
- Info. Extraction (for Annotation)
 - Conditional Random Fields (Simple Tagger)
 - Datamold / CRAM
 - Require all tokens to receive label / no junk
 - NER with Dictionary (Conditional Semi-Markov Model)
 - Whole segments receive same label – attributes can't be interrupted



Conclusion

- Annotate unstructured and ungrammatical sources
 - Don't involve users
 - Structured queries over data sources
- Future:
 - Automate entire process
 - Unsupervised RL and IE
 - Mediator gets Reference Sets
- More Info:
 - www.isi.edu/~michelso

Questions?