Mayank Kejriwal

University of Southern California

Lexicons and reference sets - II

Outline

- □ Introduction
- □ Alignment
- □ Extraction
- □ Results
- □ Discussion

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

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
Area	Phoebus	89.25	87.50	**88.28**	809.7
	Simple Tagger	92.28	81.24	86.39	
	Amilcare	74.2	78.16	76.04	
Date	Phoebus	87.45	90.62	**88.99**	751.9
	Simple Tagger	70.23	81.58	75.47	
	Amilcare	93.27	81.74	86.94	
Name	Phoebus	94.23	91.85	93.02	1873.9
	Simple Tagger	93.28	93.82	**93.54**	
	Amilcare	83.61	90.49	86.90	
Price	Phoebus	98.68	92.58	**95.53**	850.1
	Simple Tagger	75.93	85.93	80.61	
	Amilcare	89.66	82.68	85.86	
Star	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
Condition	Phoebus	91.8	84.56	**88.01**	410.3
	Simple Tagger	78.11	77.76	77.80	
	Amilcare	79.18	67.74	72.80	
Descript.	Phoebus	69.21	51.50	59.00	504.0
	Simple Tagger	62.25	79.85	**69.86**	
	Amilcare	55.14	58.46	56.39	
Issue	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
Publisher	Phoebus	83.81	95.08	**89.07**	61.1
	Simple Tagger	88.54	78.31	82.83	
	Amilcare	90.82	70.48	79.73	
Title	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	

Extraction results: Summary

	Token Level			***Hotel***	Field Level		
	Prec.	Recall	F-Mes.		Prec.	Recall	F-Mes.
Phoebus	93.60	91.79	**92.68**		87.44	85.59	**86.51**
Simple Tagger	86.49	89.13	87.79		79.19	77.23	78.20
Amilcare	86.12	86.14	86.11		85.04	78.94	81.88
	Token Level			***Comic***	Field Level		
	Prec.	Recall	F-Mes.		Prec.	Recall	F-Mes.
Phoebus	93.24	84.48	**88.64**		81.73	80.84	**81.28**
Simple Tagger	84.41	86.04	85.43		78.05	74.02	75.98
Amilcare	87.66	81.22	84.29		90.40	72.56	80.50

Results Discussion

3 attributes where Phoebus not max F-measure

- Hotel name – tiny difference
- Comic Title – low recall → lower F-measure
    - recall: missed tokens of titles not in ref. set
    - "The Incredible Hulk and Wolverine" → "The Incredible Hulk"
- Comic description
    - Simple Tagger learned internal structure of descriptions
        - High recall, low precision
    - Phoebus labels in isolation
        - Only meaningful tokens (like prop. Names) labeled
        - higher precision, lower recall → 2nd best F-measure

Outline

- □ Introduction
- □ Alignment
- □ Extraction
- □ Results
- □ Discussion

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 ref. set annotation
- Perform better at annotation than extraction
  - Consider Rec. link results as field level extraction
  - E.g. no system did well extracting comic desc.
    - +20% precision, +10% recall using rec. 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 rec. link
  - If wrong rec. link, 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