Entity Summarization of News Articles

Roi Blanco, Gianluca Demartini,
Malik Muhammad Saad Missen, Hugo Zaragoza

Internship at Yahoo!Lab Barcelona
September – November 2009

Motivation

- Going beyond document retrieval
- Finding entities relevant to a query in a document collection (e.g., Wikipedia)

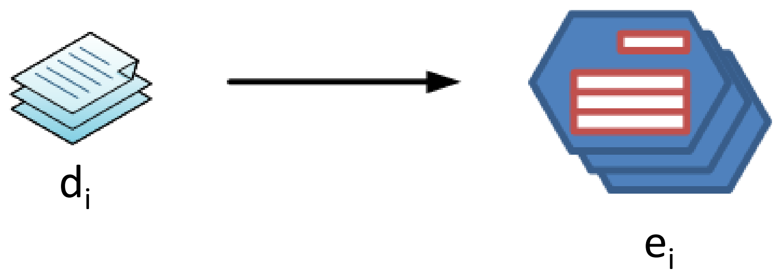
- In collections of documents over time
  - Decide about relevance at document level (Entity Summarization)
  - Analyse and exploit relevance evolution

Scenario

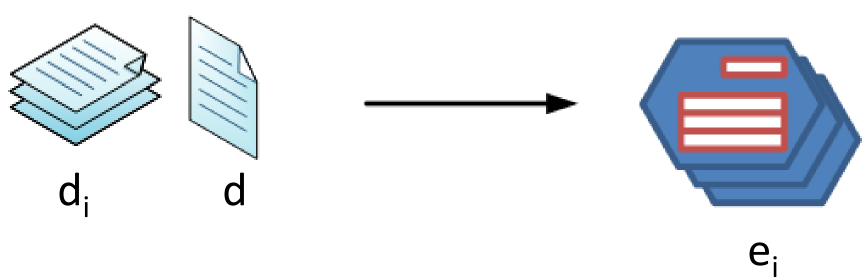
- An event
  - Charles Schulz dies
- Get Relevant Docs
- Entities
  - Peanuts, his wife, media companies, hometown, other cartoonists, ...
- Timeline of relevant news:
  - 10/1999-09/2000:
    - 11/99 cancer diagnosted
    - 12/99 he retires
    - 02/00 he dies
    - 03/00 peanuts future discussed
    - ... Honors, museums, statues, airports, ...

Tasks

- Entity Ranking (ER)
  - Find the set of entities e_i that best describe the relevant documents d_i
  - Yahoo! Correlator

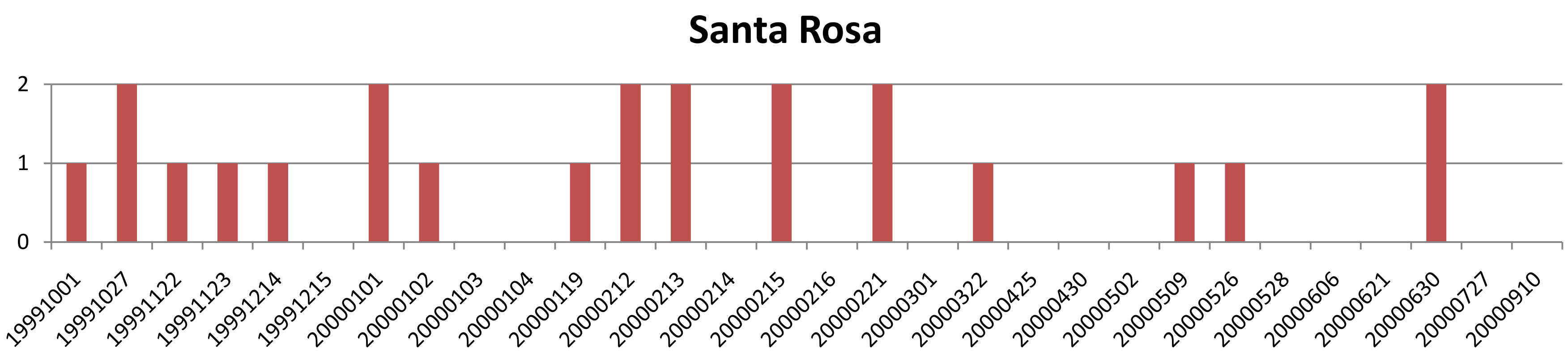


- Entity Summarization (ES)
  - Find the set of entities e_i that best describe document d wrt a query q
  - Subtask: Find e_i for d wrt a query q given history $d_i<d$



Tasks

- Entity Profiling (EP)
  - Construct temporal development of entity relevance



Outline

- Dataset
- Data analysis
- Entity Summarization
- Entity Profiles
- Conclusions

Dataset

- TREC Novelty Track 2004
  - Sentence retrieval
  - 25 event topics
  - 779 **relevant** news
- Entity annotations (7481 entities)
  - Persons (26%), Locations (10%), Organizations (57%), Products (7%)
- Relevance judgements
  - Of each entity wrt to topic in this current news
  - 21,213 judgements on 3 levels
  - Cohen's Kappa 0.59

Data Analysis

- How useful is to find relevant sentences?
  - P(e is Rel) 0.411 [0.404-0.417]
  - P(e is NotRel) 0.168 [0.163-0.173]
  - P(e is Rel | s is Rel) 0.547 [0.534-0.559]
  - Sentences:
    - 21727 total 1.46 entity occurences
    - 5122 relevant 1.88 entity occurences
    - 2122 novel 1.92 entity occurences

- How useful is to find novel sentences?
  - P(e is Rel | s is Nov) 0.510 [0.491-0.531]

Data Analysis

- How useful is looking at the past?
  - $P(e|d_1)$ 0.893 [0.881-0.905]
  - $P(e|d_{-1})$ 0.701 [0.677-0.726]

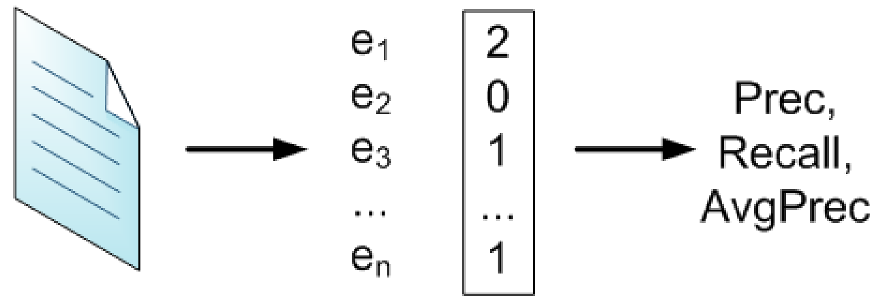
- Is useful to consider sentence co-occurence?

P(e1,e2)	**Relevant**	**Related**	**NotRelevant**	**NotAnEntity**
Relevant	0.24	0.08	0.03	0.07
Related		0.07	0.03	0.03
NotRelevant			0.07	0.05
NotAnEntity				0.04

Outline

- Dataset
- Data analysis
- **Entity Summarization**
  - Local Features
  - History Features
- Entity Profiles
- Conclusions

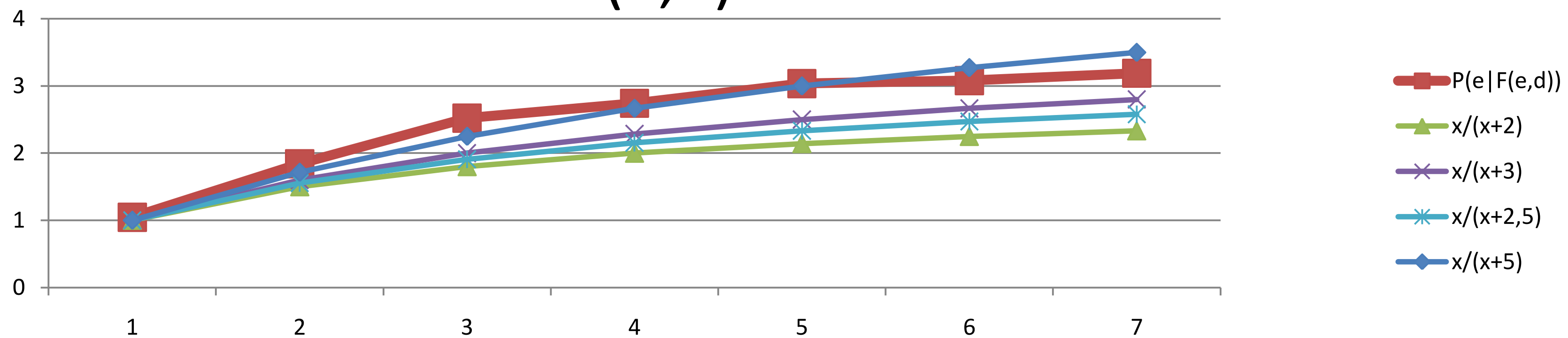
Entity Summarization



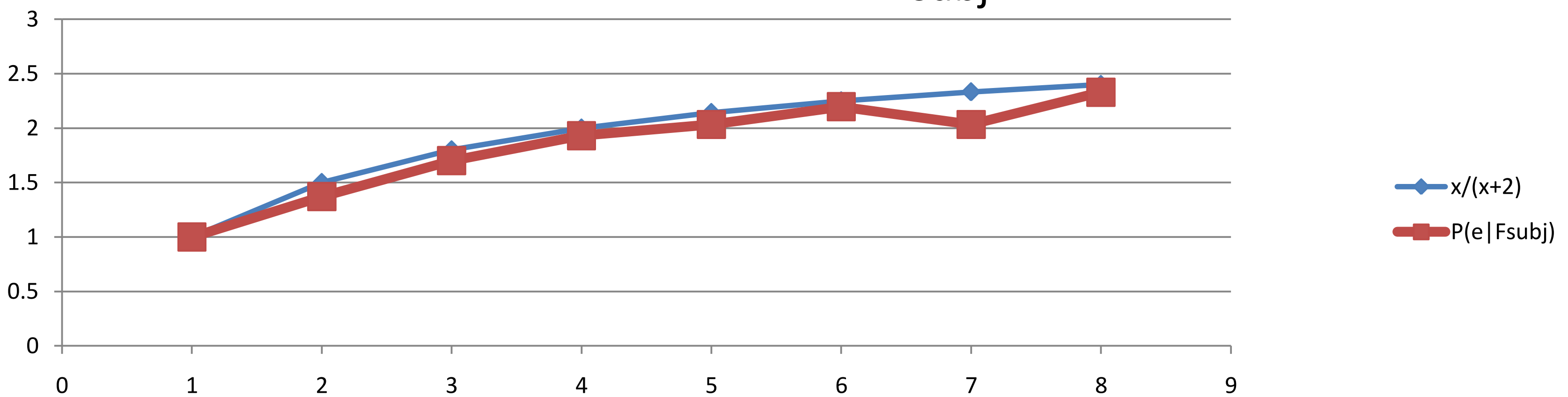
- Evaluation
  - P3, P5, AvgPrec
  - Ties aware measures [McSherry and Najork, ECIR08]
- Paired t-test
  - ** $p<<0.01$
  - * $p<0.05$
- Related considered NonRelevant

Local Features

- Looking at the document
  - Occurences of e: F(e,d)

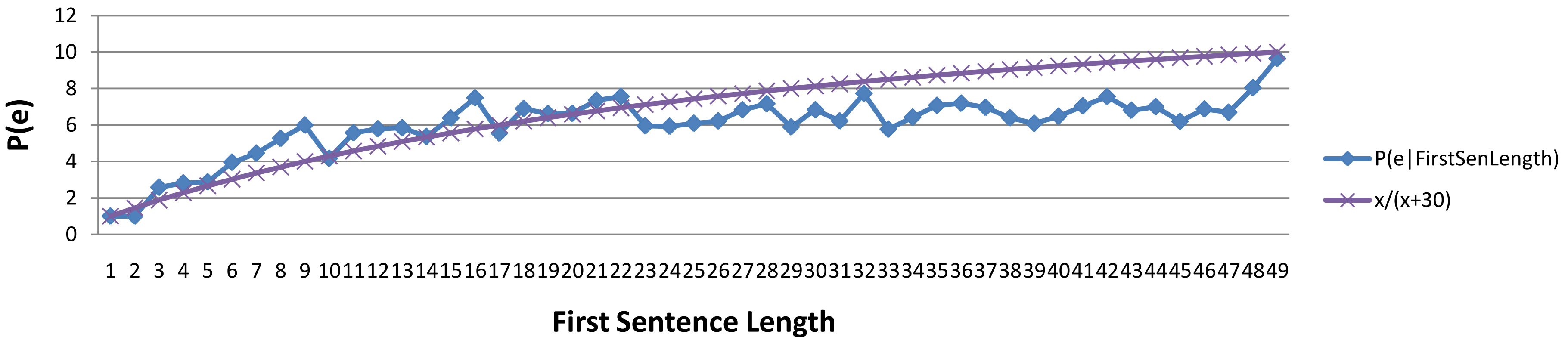


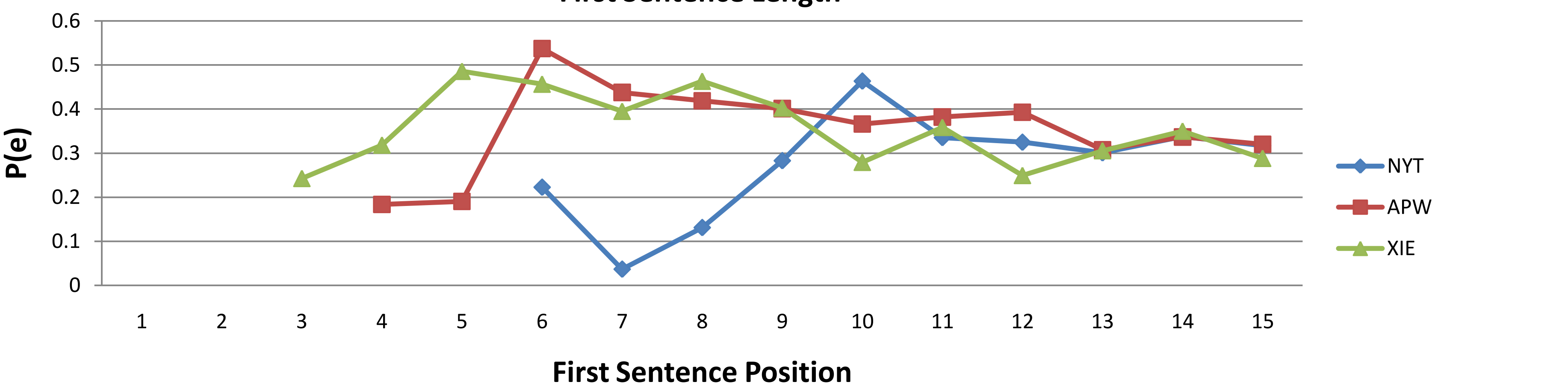
  - Occurences of e as subject: $F_{subj}(e,d)$



Local Features

- Look at the position of e in the document
  - Lenght of the first sentence where e appears
  - Position of the first sentence where e appears





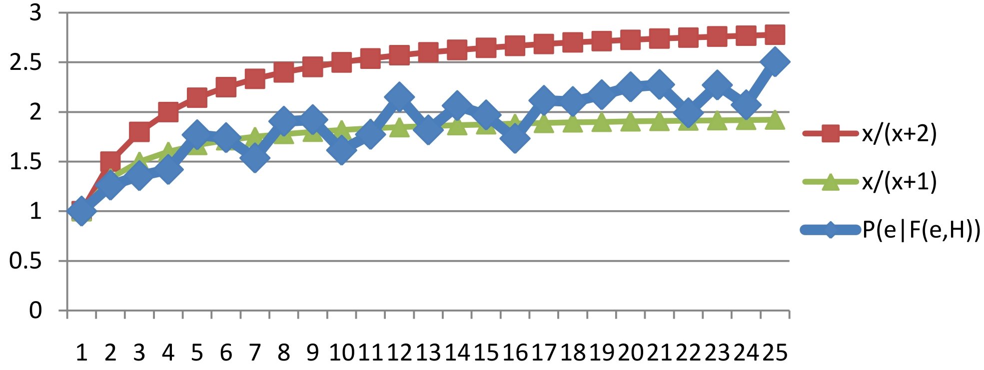
Local Features

Feature	P3	P5	MAP
F(e,d)	**.65**	**.56**	**.60**
FirstSenLen	.37	.36	.45
FirstSenPos	.31	.31	.43
F_{subj}	.49	.44	.50
AvgBM25s	.27	.30	.41
SumBM25s	.50	.44	.52

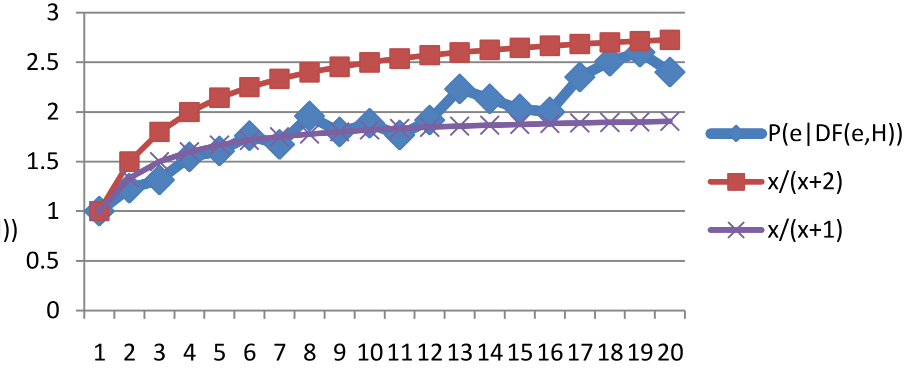
Feature	P3	P5	MAP
All Tied	.34	.34	.42

Entity Summarization

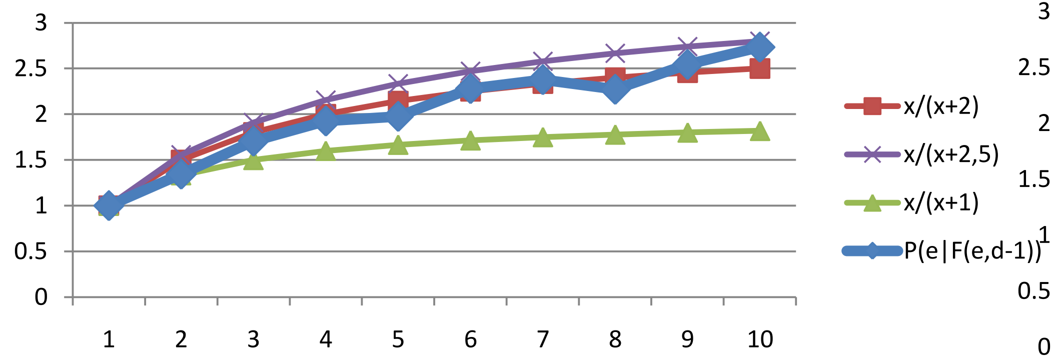
- Look at previous documents
  - Entity occurences so far **F(e,H)**
  - Docs where the entity appeared so far **DF(e,H)**
  - Entity occurences in the previous doc $\mathbf{F(e,d_{-1})}$
  - Frequency of entity the first time? $\mathbf{F(e,d_1)}$
  - Number of other entities with which the entity co-occured so far **CoOcc(e,H)**



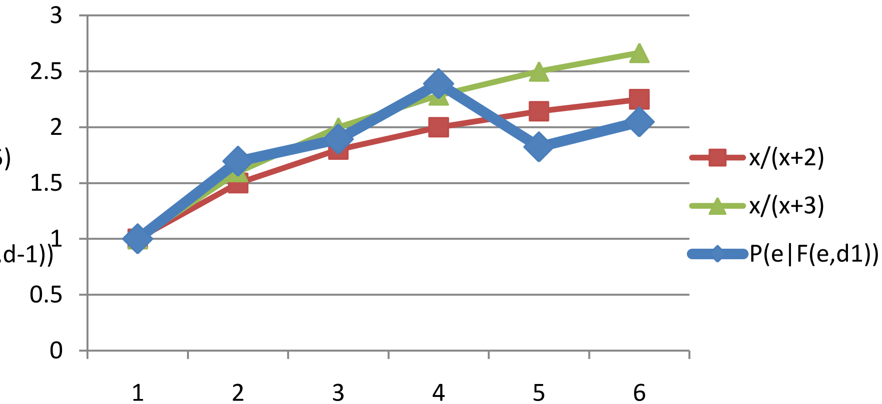
0
0.5
1
1.5
2
2.5
3
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
x/(x+2)
x/(x+1)
P(e|F(e,H))



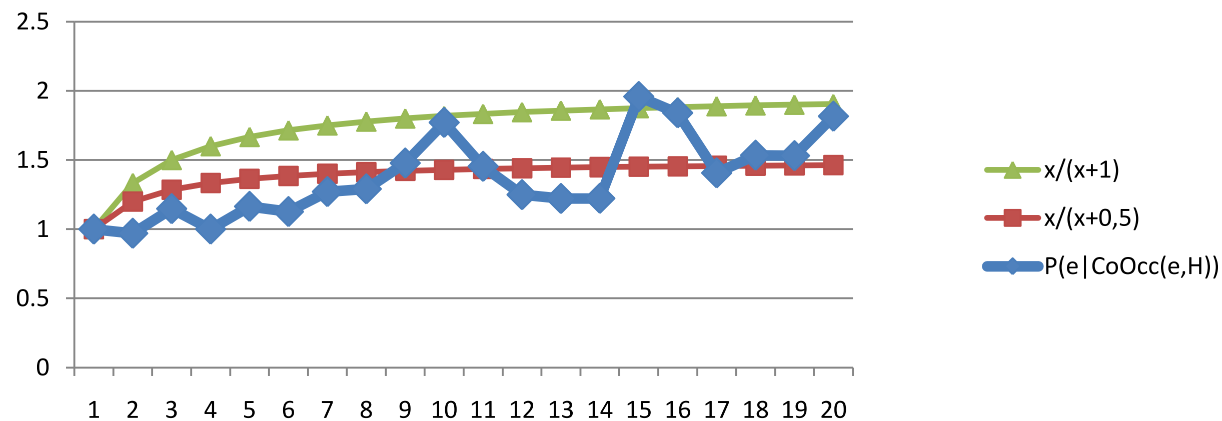
0
0.5
1
1.5
2
2.5
3
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
P(e|DF(e,H))
x/(x+2)
x/(x+1)



0
0.5
1
1.5
2
2.5
3
1 2 3 4 5 6 7 8 9 10
x/(x+2)
x/(x+2,5)
x/(x+1)
P(e|F(e,d-1))



0
0.5
1
1.5
2
2.5
3
1 2 3 4 5 6
x/(x+2)
x/(x+3)
P(e|F(e,d1))



0
0.5
1
1.5
2
2.5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
x/(x+1)
x/(x+0,5)
P(e|CoOcc(e,H))

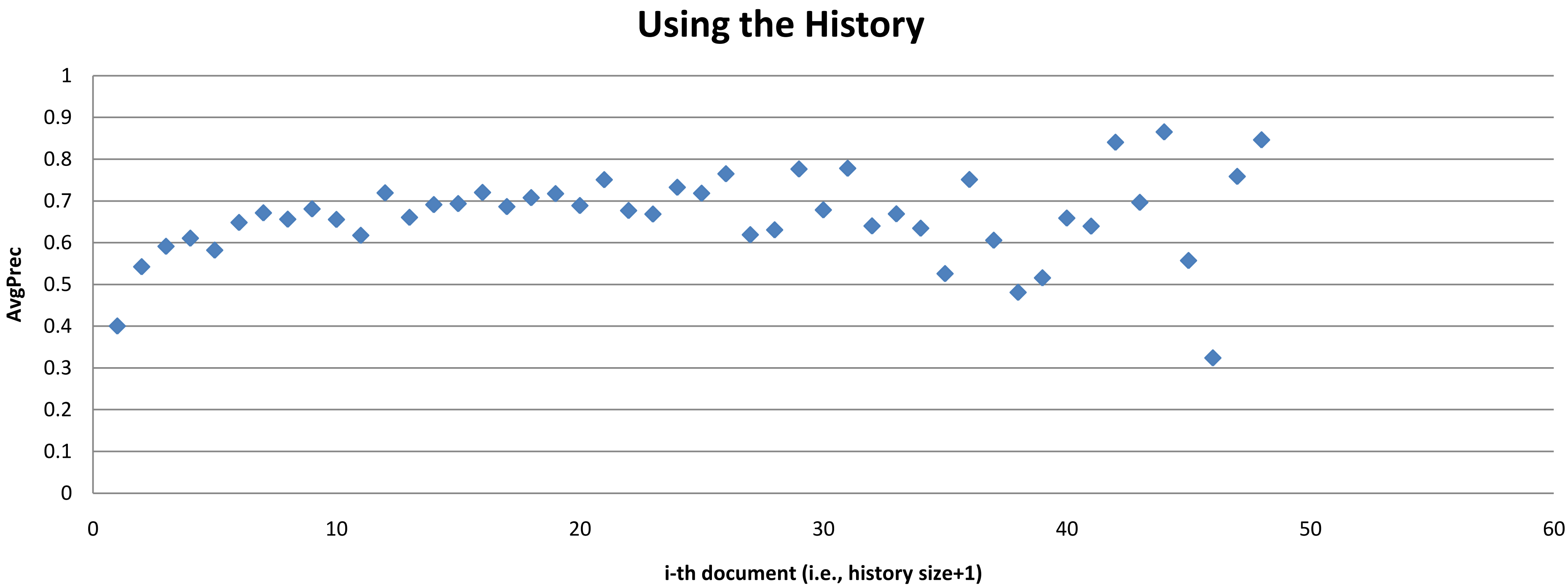
History Features

Feature	P3	P5	MAP
F(e,d)	.65	.56	.60
$F(e,d_1)$	.58	.53	.56
$F(e,d_{-1})$	.64	.56	.62*
F(e,H)	**.66**	**.59****	**.66****
CoOcc(e,H)	.62	.57	.65**
DF(e,H)	.63	.57*	.65**

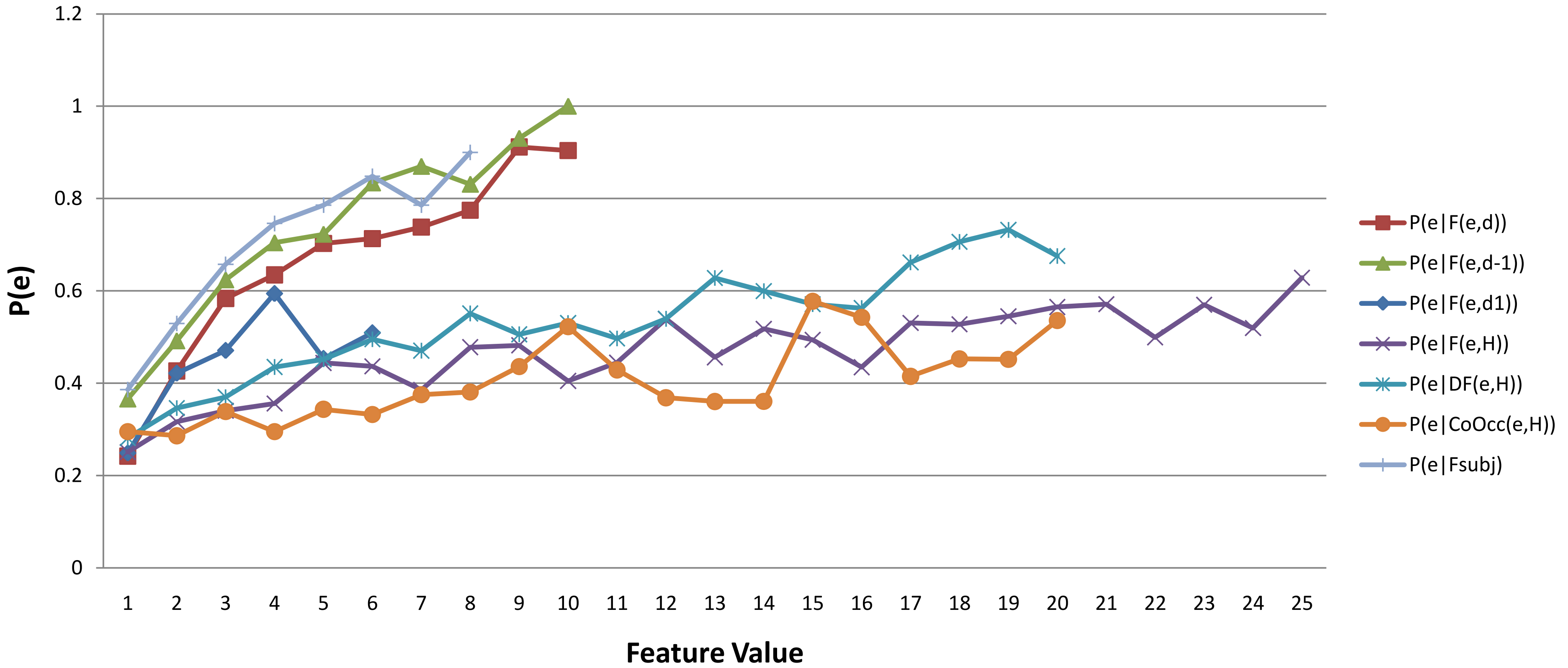
- We also tried
  - Weight history features with doc length
  - Weight history features with BM25

Using the History

- Conclusion
  - Evidence from past documents is very important
  - Effectiveness should improve over time (run F(e,H))

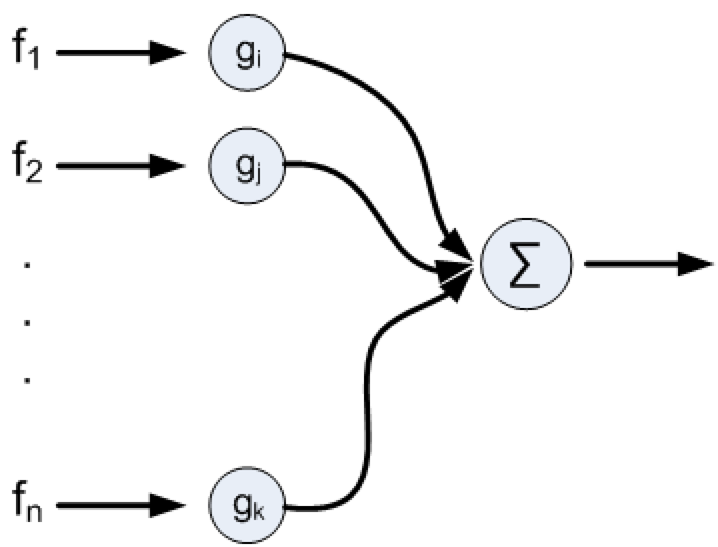


Comparing Features



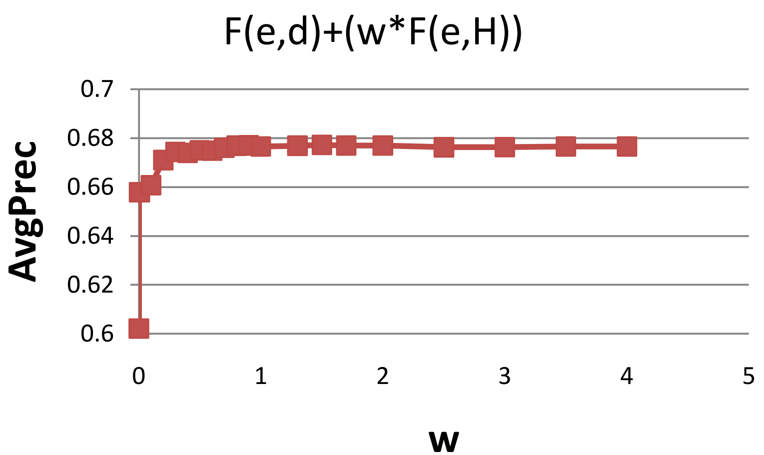
Feature Combinations

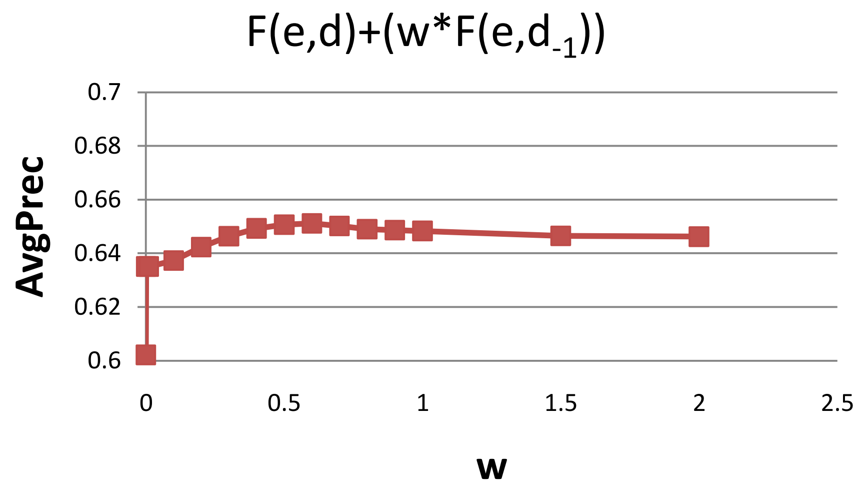
$$score(e, \vec{f}) = \sum_{i=1}^{n} w_i g(f_i, \Theta_i) \qquad g(x,t) = \frac{x}{x+t}$$

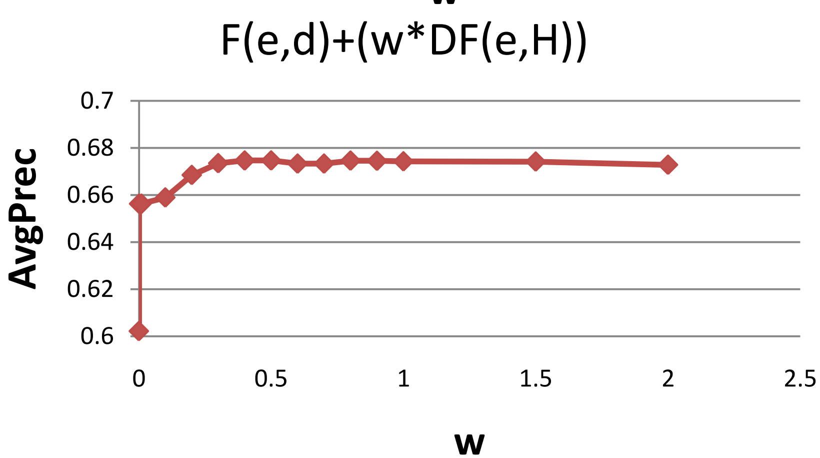


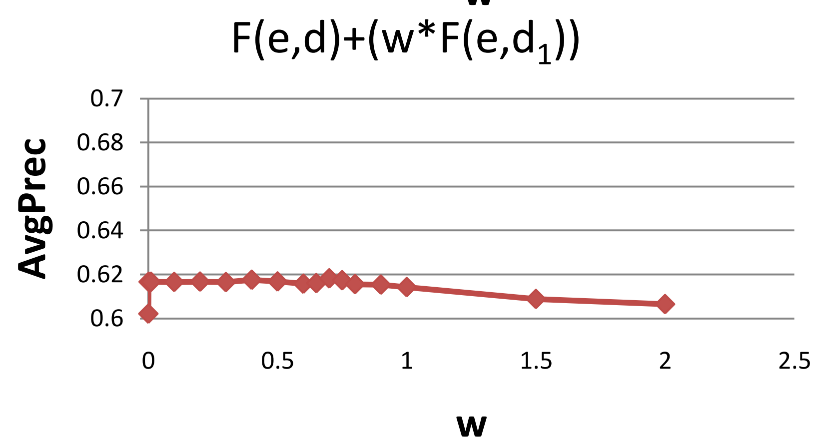
Combining 2 Features

$$score(e, F_1, F_2) = \left(\frac{F_1}{F_1 + t_1}\right) + w\left(\frac{F_2}{F_2 + t_2}\right)$$

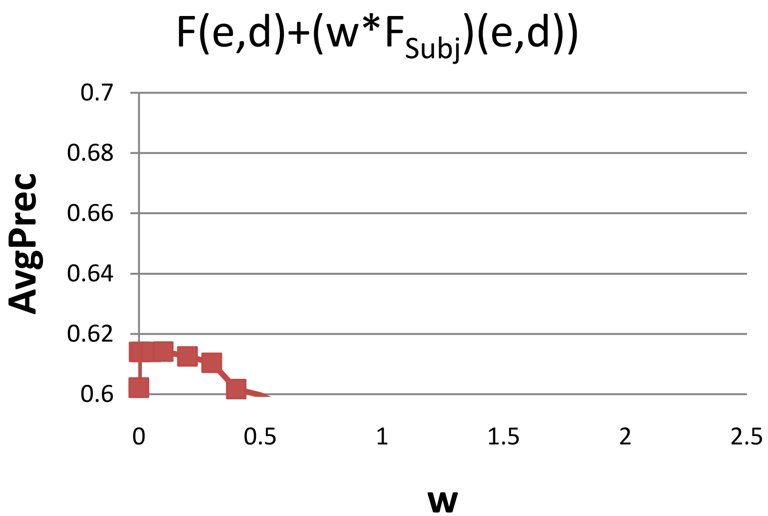


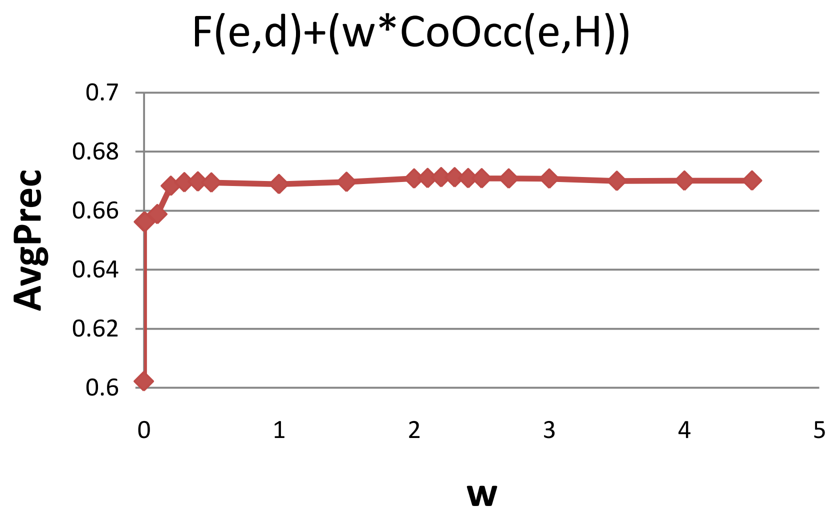


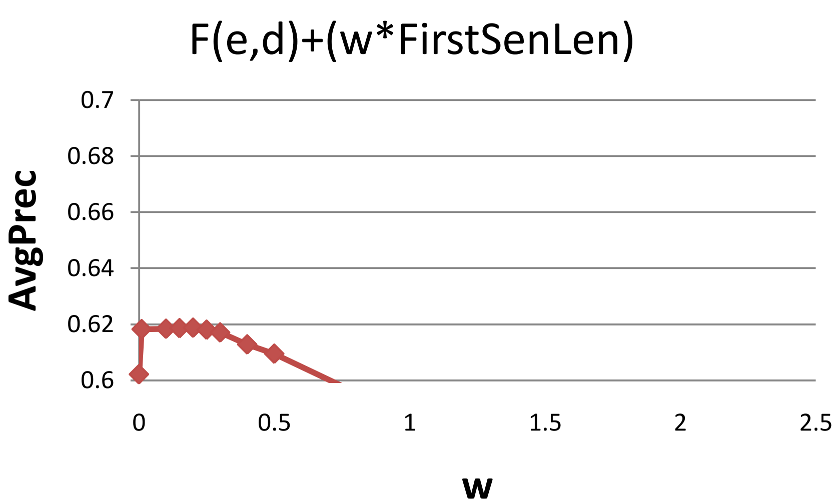


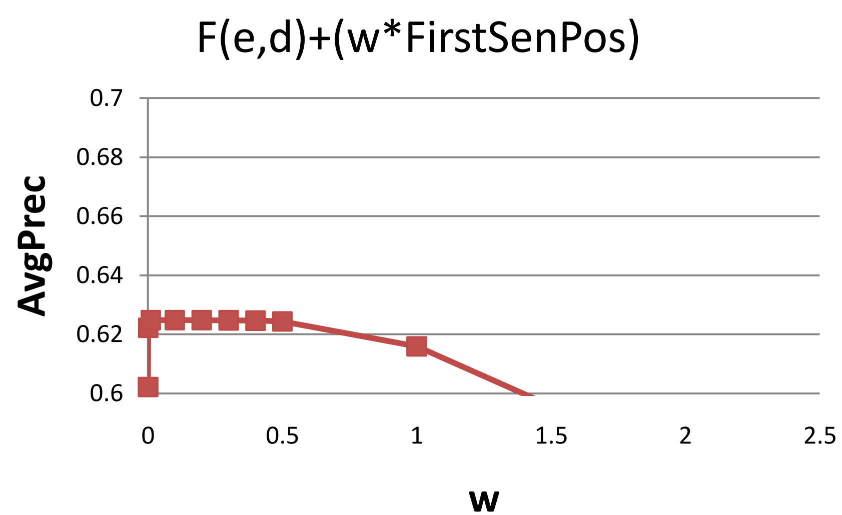


Combining 2 Features









Combining 2 features

Feature	Function g	W	(F(e,d)+wF) AvgPrec
F(e,d)		-	.60
F(e,H)			.66
FirstSenLen	x/(x+30)	0.2	.62**
FirstSenPos	x/(x+2)	0.1	.62**
F_{subj}	x/(x+2)	0.1	.61**
$F(e,d_1)$	x/(x+3)	0.7	.62**
$F(e,d_{-1})$	x/(x+2)	0.6	.65**
F(e,H)	x/(x+1)	1.5	**.68****$^{++}$
CoOcc(e,H)	x/(x+0.5)	2.2	.67**$^{+}$
DF(e,H)	x/(x+1)	0.5	.67**$^{++}$

Combining 3 features

$$score(e, F_1, F_2, F_3) = (\frac{F_1}{F_1 + t_1}) + w_1(\frac{F_2}{F_2 + t_2}) + w_2(\frac{F_3}{F_3 + t_3})$$

- Optimizing w_1, w_2

F_1	F_2	F_3	w_1, w_2	AvgPrec
F(e,d)	F(e,d_{-1})	F(e,H)	0.4,1.0	.68
F(e,d)	CoOcc(e,H)	F(e,H)	0.12,1.84	.68
F(e,d)	CoOcc(e,H)	FistSenLen	2.1,0.01	.67
F(e,d)	CoOcc(e,H)	F(e,d_1)	2.2,0	.67

- Optimizing t_1, t_2, t_3

F_1	F_2	F_3	t_1, t_2, t_3	AvgPrec
F(e,d)	F(e,d_{-1})	F(e,H)	5.9, 6.9, 13.8	**.69****

Combining Features with ML

- Logistic Regression for ranking entities
- 5-folds cross validation on 25 topics
- Similar results for combinations of 2 features

Local Doc Features
F(e,d)
FirstSenLen
FirstSenPos
F_{subj}
AvgBM25s
SumBM25s

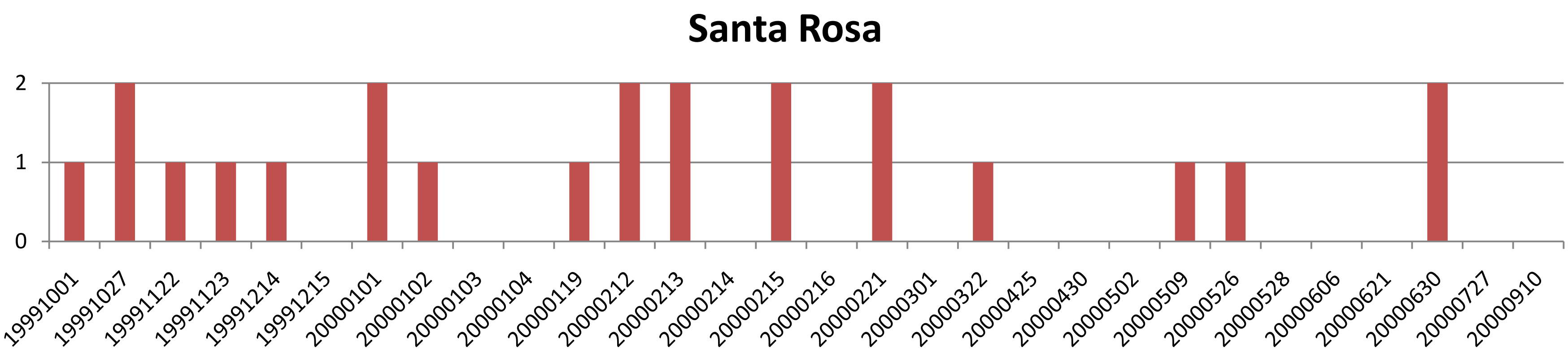
History Features
$F(e,d_1)$
$F(e,d_{-1})$
F(e,H)
CoOcc(e,H)
DF(e,H)

Features	P3	P5	AvgPrec
F(e,d)	.65	.56	.60
Local	.65	.56	.62
History	.66	.60	.67
All	.69	.62	.68

Outline

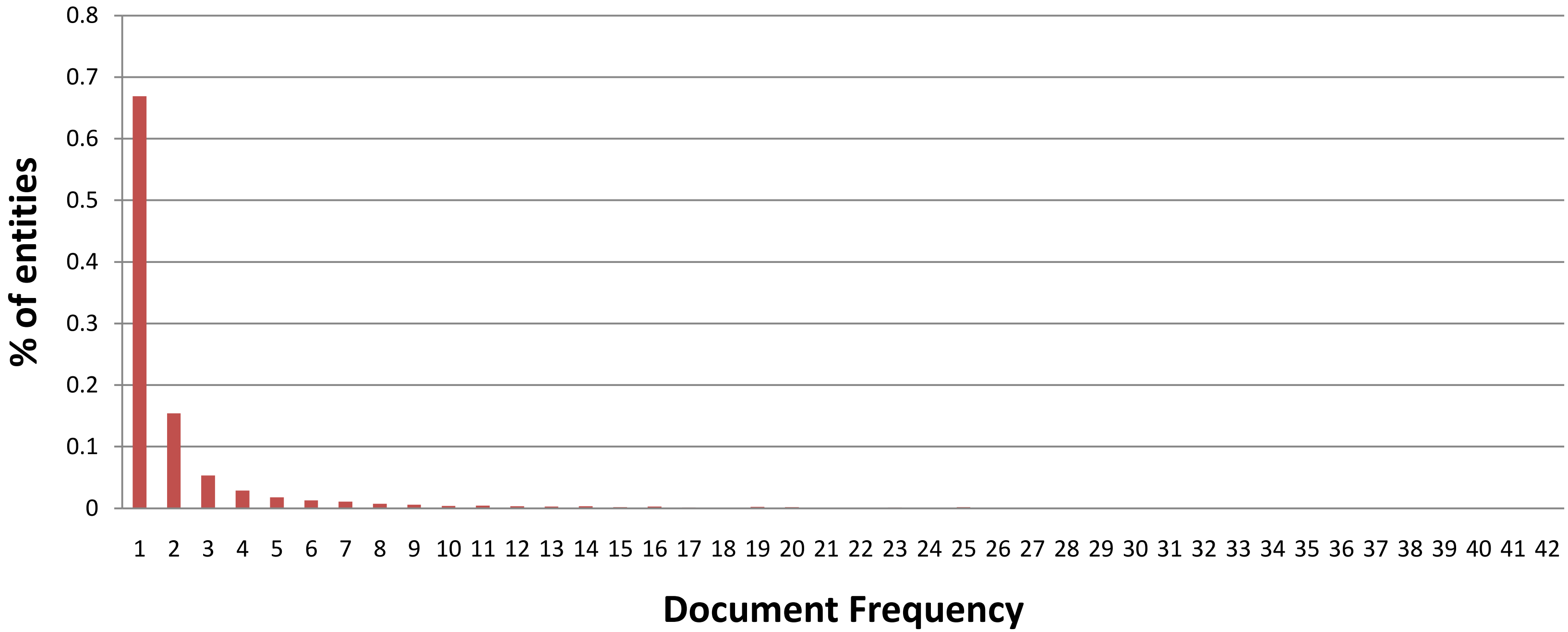
- Dataset
- Data analysis
- Entity Summarization
- **Entity Profiles**
- Conclusions

Entity Profiles

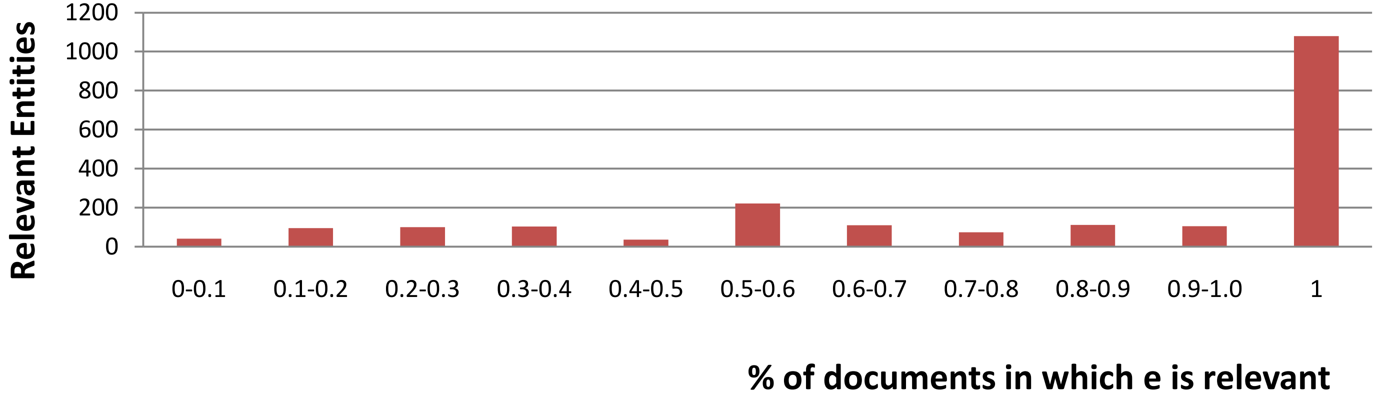


- There are lots of entities in a story
  - 31 judged docs per topic
  - 27 judged entities per doc
- For which entities we should build profiles?

- 67% of entities appear only in 1 document



- Relevant entities stay relevant all the time



- We pick 708 entities
  - That are relevant at least one day
  - That do not have always the same judgement

Future prediction

- Query: Will entity e be relevant in future?
- Predict appearence (and relevance) of an entity e in future documents given that
  - e has appeared in the past (as Relevant)
  - e does not appear today
  - 7% of entities appear
    - as at least twice as Relevant
    - with a gap in their profiles

Future prediction

- Goal: Extend summary with entities not present in the current doc
- Possible approach:
  - rank e from past docs
  - filter e in current d
  - evaluate with ground truth on future

Conclusions

- Defined new tasks: ES, EP
- Constructed evaluation benchmark
- Entity summarization
  - Investigated some features and combinations
  - Information from the past helps most
  - Obtain 15% improvement over F(e,d)
- Entity profiling
  - How to select interesting entities

Entity Profiles

- Evaluation setting
  - 708 entities
  - Average over all ON/OFF decisions (some entities may have more decisions)
  - Skip non judged (entity,date) pairs
  - Related considered ON-OFF
  - Skip entities which have never been relevant!
  - tp=ON,Rel tn=OFF,NonRel fp=ON, NonRel fn=OFF,Rel

Entity Profiles

Algo	P=tp/(tp+fp)	R=tp/(tp+fn)
Always ON	**.84 - .56**	**1 - 1**
Always OFF	**0 - 0**	**0 - 0**
Freq-based	**.84 - .57**	**.77 - .77**
withPast (t/5)	**.85 - .60**	**.51 - .54**
withPast (t/3)	**.85 - .61**	**.45 - .48**
withPast (t/2)	**.85 - .62**	**.39 - .42**
with Past (t)	**.89 - .68**	**.29 - .33**