Selective Sampling for Information Extraction with a Committee of Classifiers

Evaluating Machine Learning for Information Extraction, Track 2

Ben Hachey, Markus Becker, Claire Grover & Ewan Klein
University of Edinburgh

Overview

- **Introduction**
  - Approach & Results
- Discussion
  - Alternative Selection Metrics
  - Costing Active Learning
  - Error Analysis
- Conclusions

Approaches to Active Learning

- Uncertainty Sampling (Cohn et al., 1995)

  Usefulness ≈ uncertainty of single learner
  - Confidence: Label examples for which classifier is the least confident
  - Entropy: Label examples for which output distribution from classifier has highest entropy
- Query by Committee (Seung et al., 1992)

  Usefulness ≈ disagreement of committee of learners
  - Vote entropy: disagreement between winners
  - KL-divergence: distance between class output distributions
  - F-score: distance between tag structures

Committee

- Creating a Committee
  - Bagging or randomly perturbing event counts, random feature subspaces (Abe and Mamitsuka, 1998; Argamon-Engelson and Dagan, 1999; Chawla 2005)
    - Automatic, but not ensured diversity…
  - Hand-crafted feature split (Osborne & Baldridge, 2004)
    - Can ensure diversity
    - Can ensure some level of independence
- We use a hand crafted feature split with a maximum entropy Markov model classifier (Klein et al., 2003; Finkel et al., 2005)

Feature Split

Feature Set 1	
Word Features	*w_i, w_{i-1}, w_{i+1}*
	Disjunction of 5 prev words
	Disjunction of 5 next words
Word Shape	*$shape_i$, $shape_{i-1}$, $shape_{i+1}$*
	$shape_i + shape_{i+1}$
	$shape_i + shape_{i-1} + shape_{i+1}$
Prev NE	*NE_{i-1}, $NE_{i-2} + NE_{i-1}$*
	$NE_{i-3} + NE_{i-2} + NE_{i-1}$
Prev NE + Word	*$NE_{i-1} + w_i$*
Prev NE + shape	*$NE_{i-1} + shape_i$*
	$NE_{i-1} + shape_{i+1}$
	$NE_{i-1} + shape_{i-1} + shape_i$
	$NE_{i-2} + NE_{i-1} + shape_{i-2} + shape_{i-1} + shape_i$
Position	*Document Position*

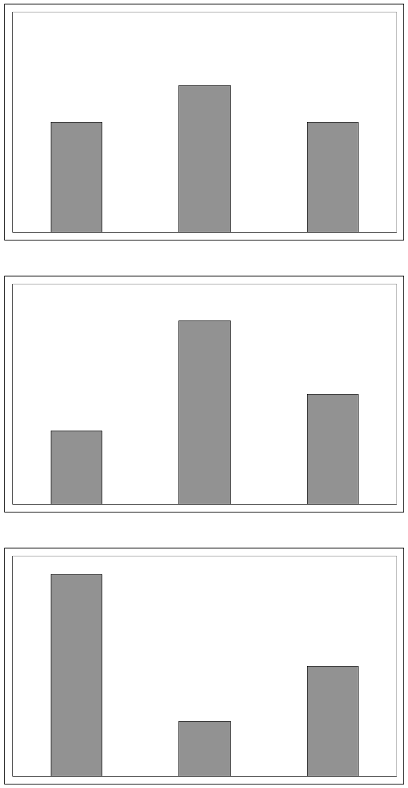
Feature Set 2	
TnT POS tags	*POS_i, POS_{i-1}, POS_{i+1}*
Prev NE	*NE_{i-1}, $NE_{i-2} + NE_{i-1}$*
Prev NE + POS	*$NE_{i-1} + POS_{i-1} + POS_i$*
	$NE_{i-2} + NE_{i-1} + POS_{i-2} + POS_{i-1} + POS_i$
Occurrence Patterns	*Capture multiple references to NEs*

Words, Word shapes, Document position	Parts-of-speech, Occurrence patterns of proper nouns

KL-divergence (McCallum & Nigam, 1998)

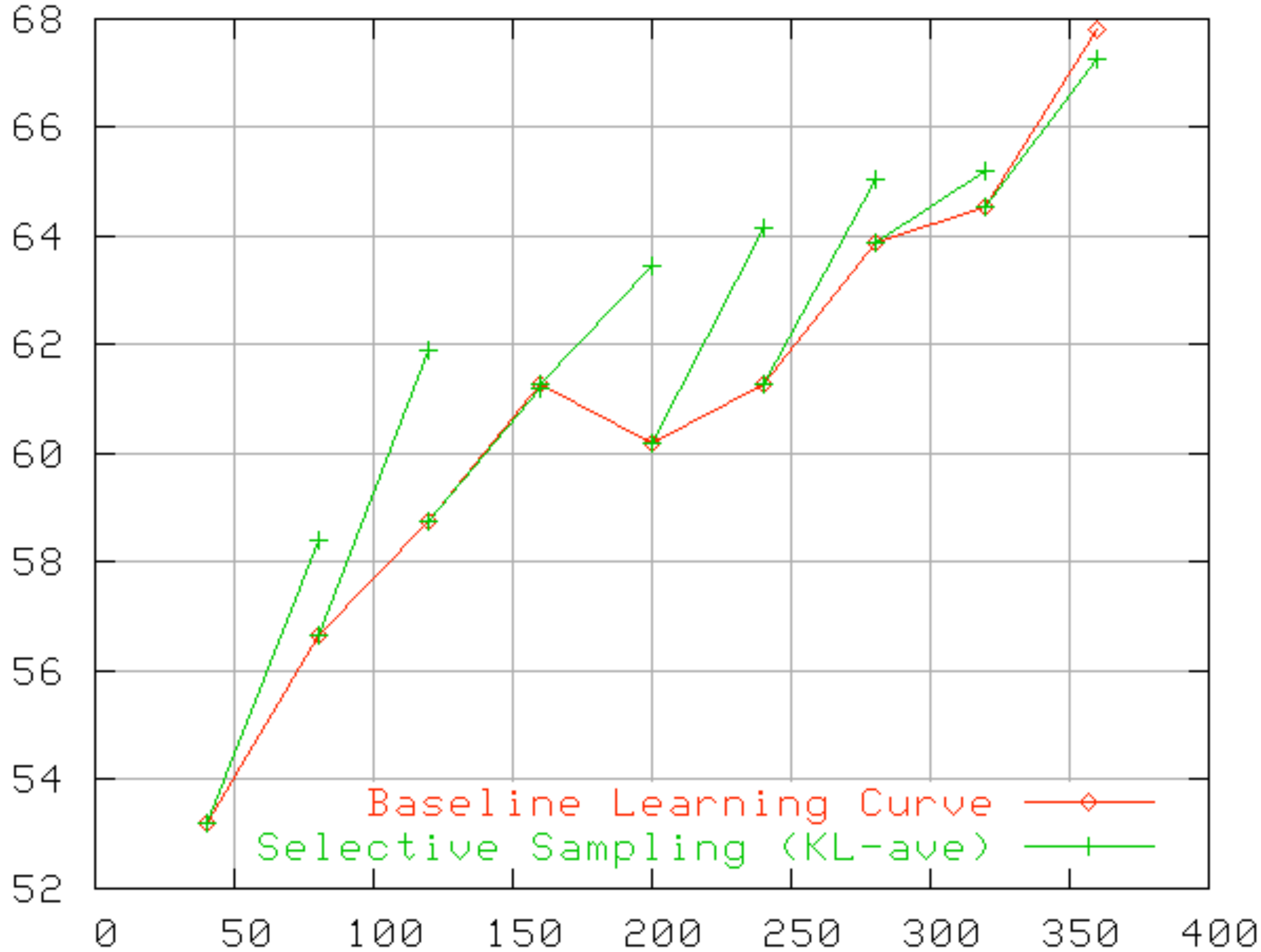
- Quantifies degree of disagreement between distributions:

$$D(p \,||\, q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$



- Document-level
  - Average

Evaluation Results



Discussion

- Best average improvement over baseline learning curve:

  1.3 points f-score
- Average % improvement:

  2.1% f-score

- Absolute scores middle of the pack

Overview

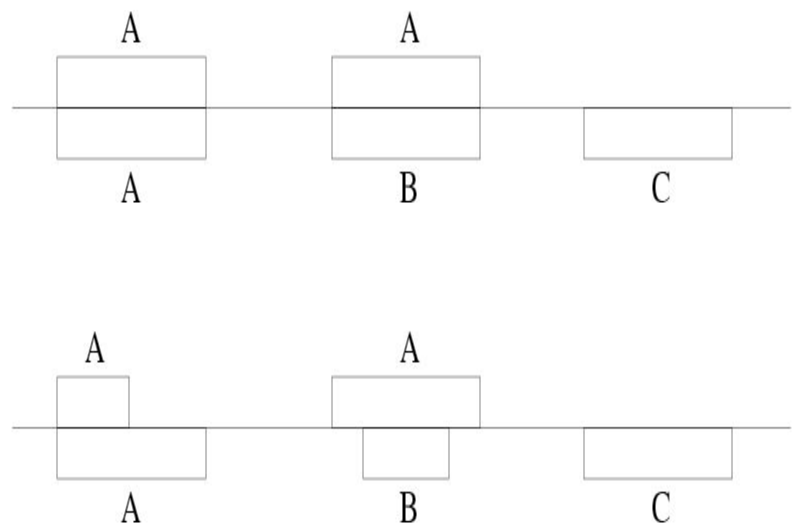
- Introduction
  - Approach & Results
- **Discussion**
  - Alternative Selection Metrics
  - Costing Active Learning
  - Error Analysis
- Conclusions

Other Selection Metrics

- KL-max
  – Maximum per-token KL-divergence
- F-complement

  (Ngai & Yarowsky, 2000)

  – Structural comparison between analyses
  – Pairwise f-score between phrase assignments:



$$f_{comp} = 1 - F(A_1(s), A_2(s))$$

Related Work: BioNER

- NER-annotated sub-set of GENIA corpus (Kim et al., 2003)
  - Bio-medical abstracts
  - 5 entities:
    `DNA, RNA, cell line, cell type, protein`
- Used 12,500 sentences for simulated AL experiments
  - Seed: 500
  - Pool: 10,000
  - Test: 2,000

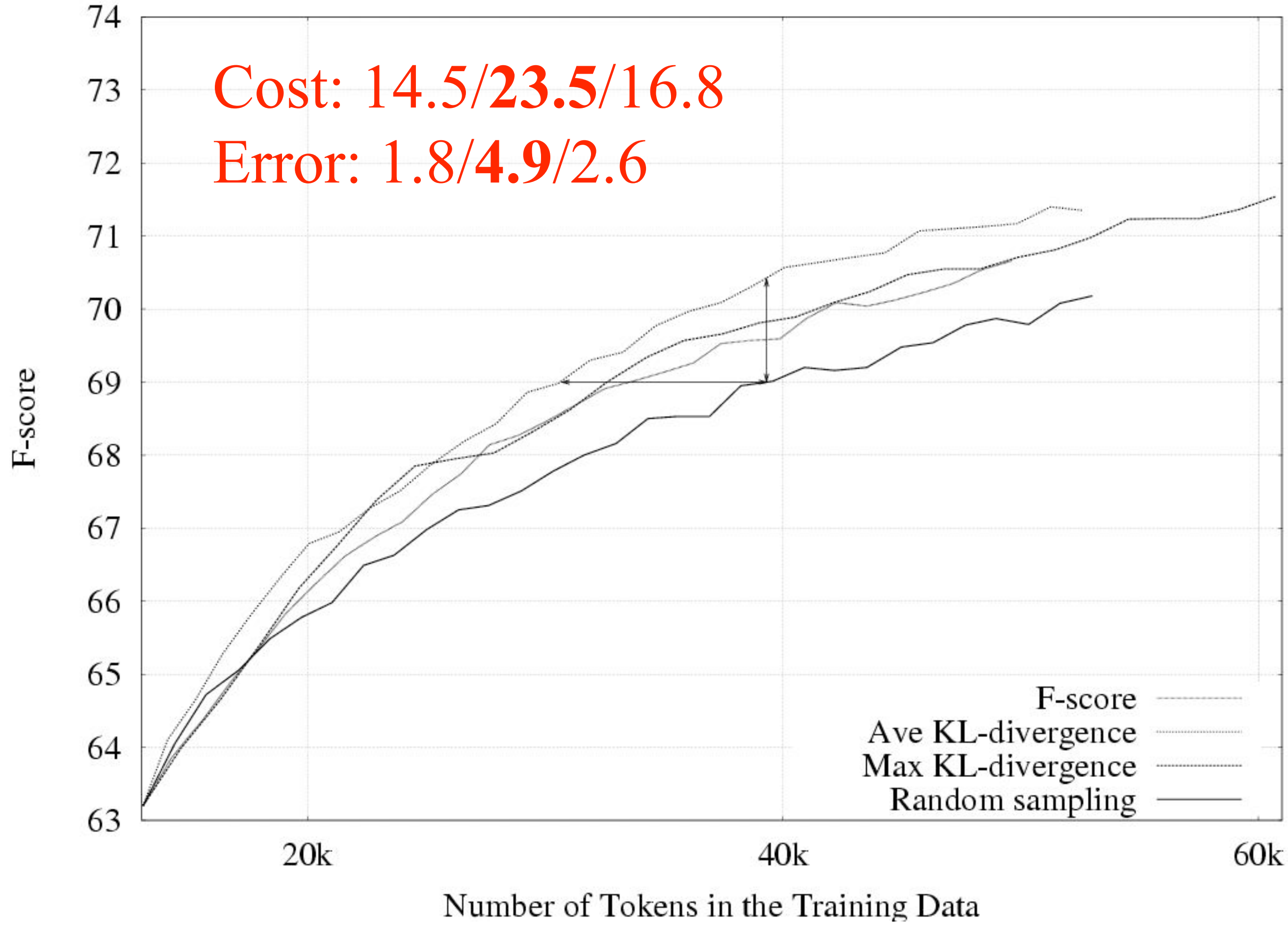
Costing Active Learning

- Want to compare reduction in cost (annotator effort & pay)

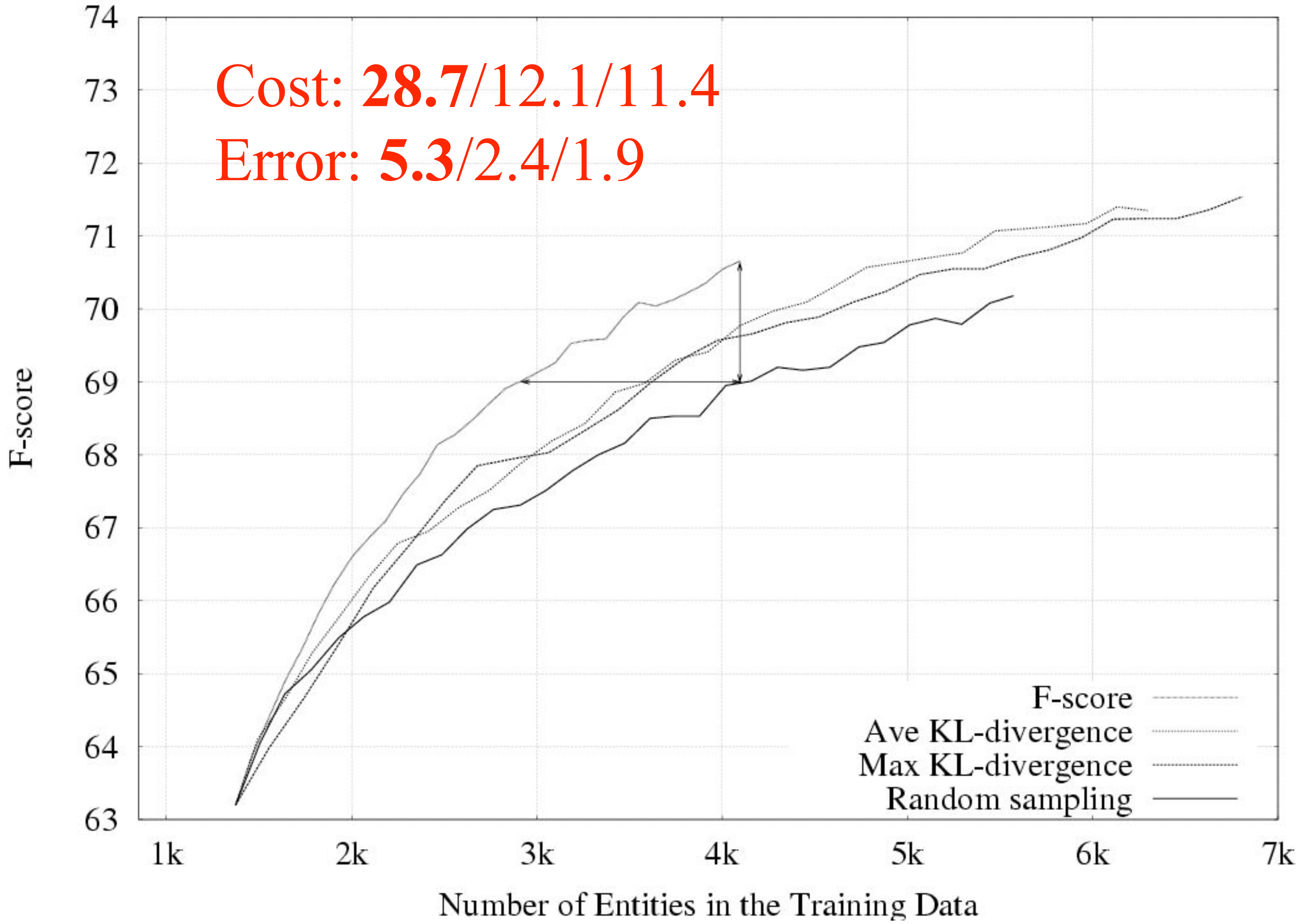
- Plot results with several different cost metrics
  - # Sentence, # Tokens, # Entities

Simulation Results: Sentences

Simulation Results: Tokens



Simulation Results: Entities



Costing AL Revisited (BioNLP data)

Metric	Tokens	Entities	Ent/Tok
Random	26.7 (0.8)	2.8 (0.1)	10.5 %
F-comp	25.8 (2.4)	2.2 (0.7)	8.5 %
MaxKL	30.9 (1.5)	3.3 (0.2)	10.7 %
AveKL	27.1 (1.8)	3.3 (0.2)	12.2 %

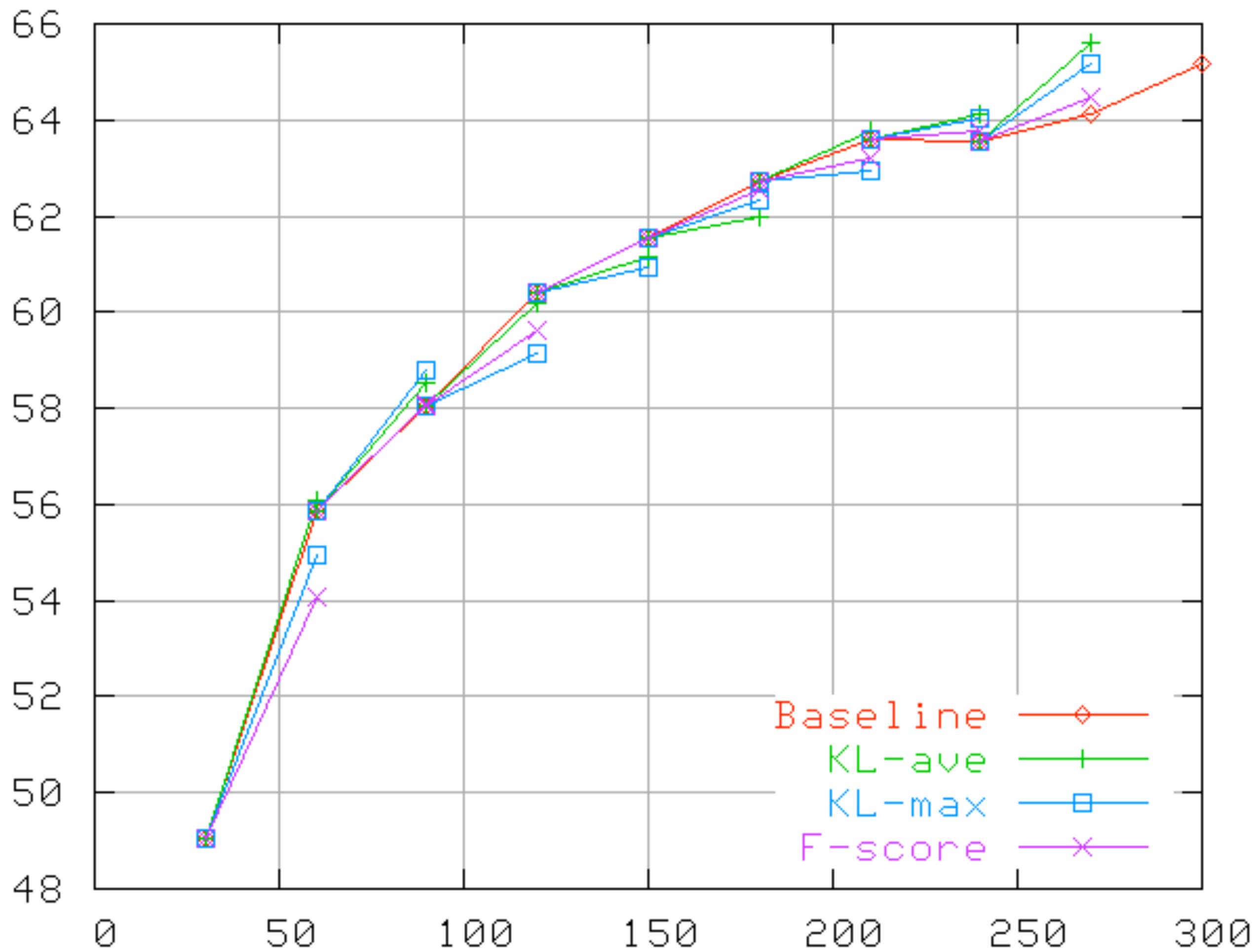
- Averaged KL does not have a significant effect on sentence length

  → *Expect shorter per sent annotation times.*

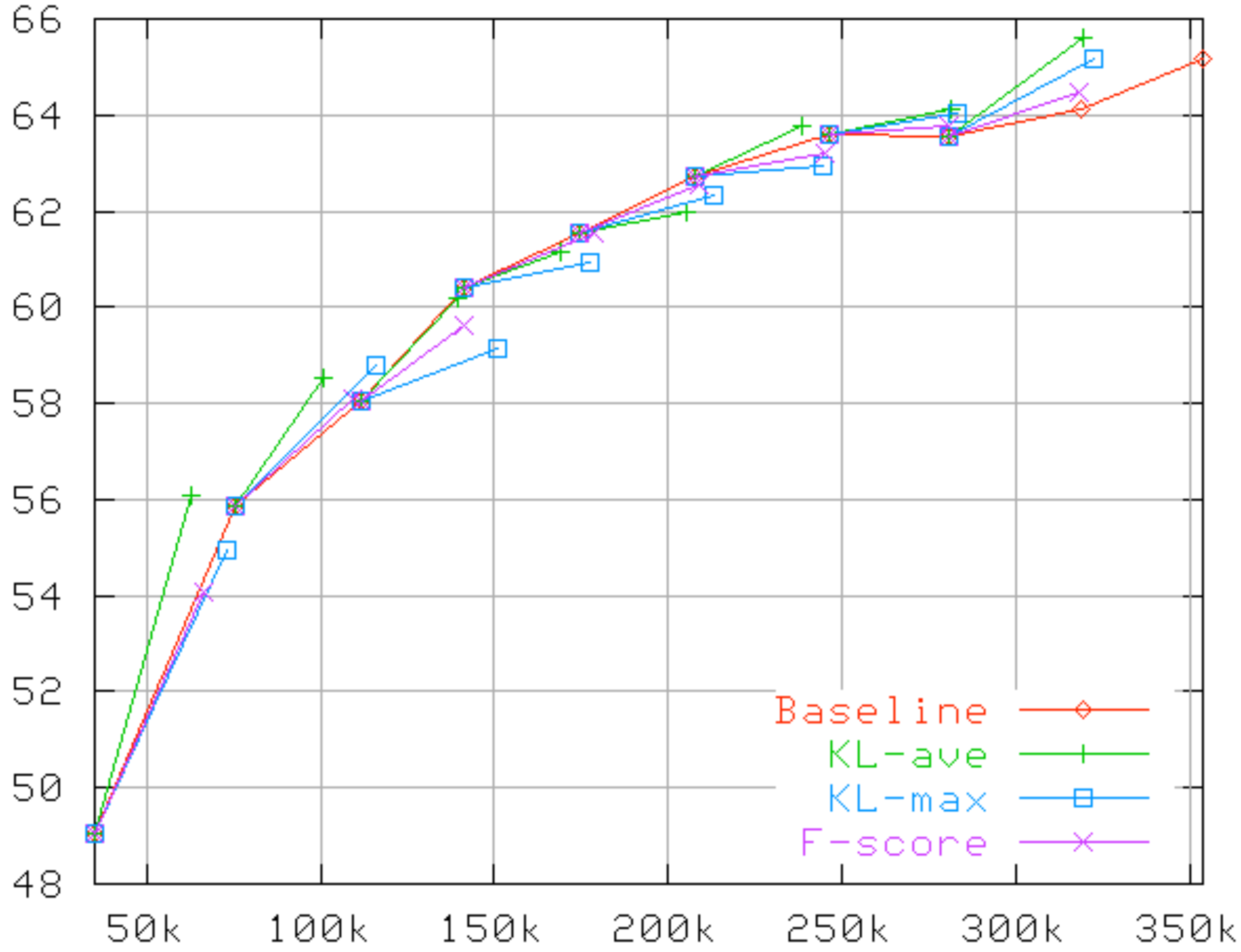
- Relatively high concentration of entities

  → *Expect more positive examples for learning.*

Document Cost Metric (Dev)



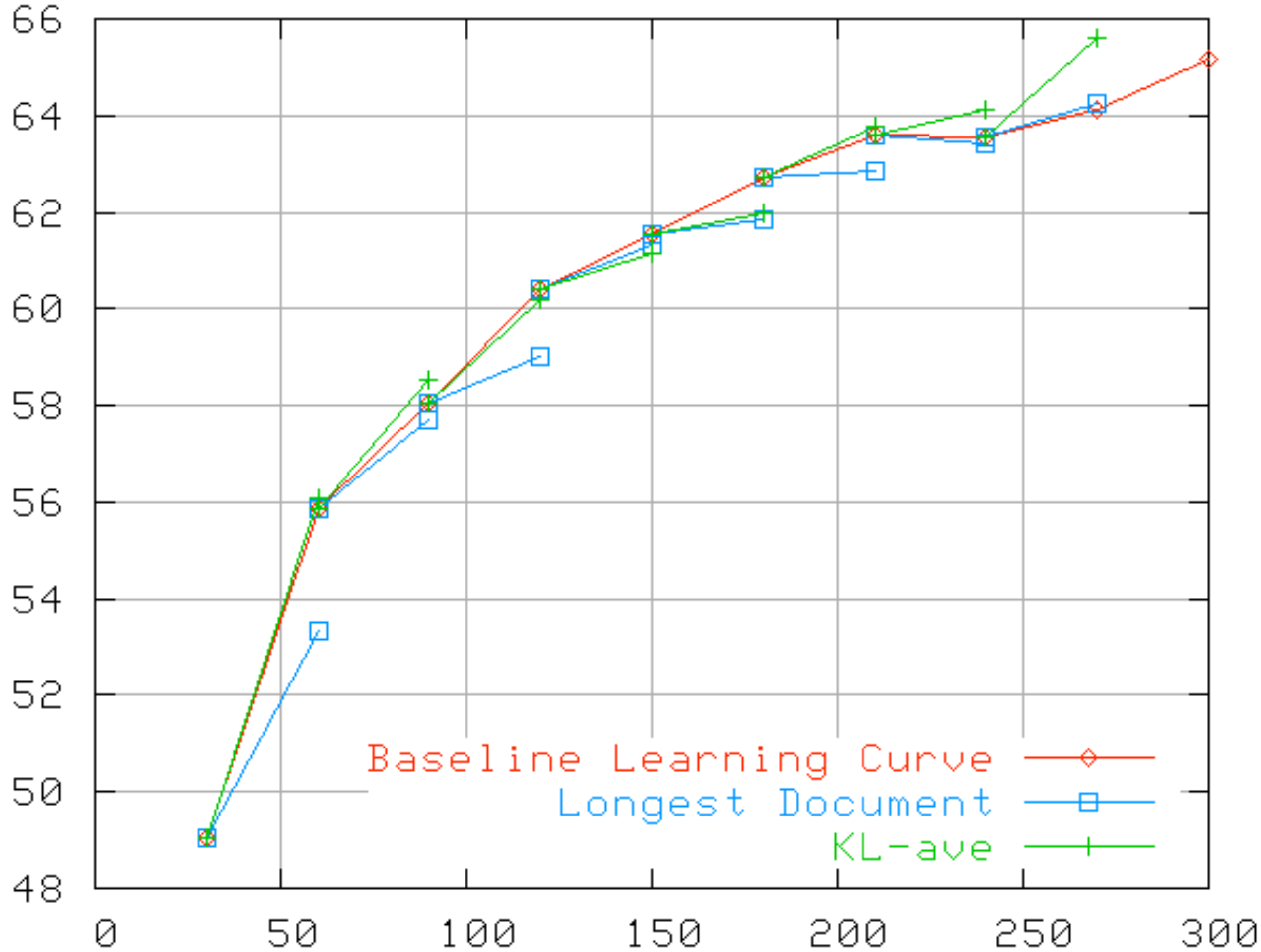
Token Cost Metric (Dev)



Discussion

- Difficult to do comparison between metrics
  - Document unit cost not necessarily realistic estimate real cost
- Suggestion for future evaluation:
  - Use corpus with measure of annotation cost at some level (document, sentence, token)

Longest Document Baseline



Confusion Matrix

- Token-level
- B-, I- removed
- Random Baseline
  - Trained on 320 documents
- Selective Sampling
  - Trained on 280+40 documents

random	O	wshm	wsnm	cfnm	wsac	wslo	cfac	wsdt	wssdt	wsndt	wscdt	cfhm
O	94.82	0.37	0.14	0.07	0.04	0.04	0.05	0.04	0.02	0.01	0.01	0.03
wshm	0.35	0.86	0	0	0	0	0	0	0	0	0	0.14
wsnm	0.34	0	0.64	0	0	0	0	0	0	0	0	0
cfnm	0.09	0	0.01	0.2	0	0	0	0	0	0	0	0
wsac	0.1	0	0	0	0.19	0	0.04	0	0	0	0	0
wslo	0.16	0	0	0	0	0.19	0	0	0	0	0	0
cfac	0.05	0	0	0	0.03	0	0.15	0	0	0	0	0
wsdt	0.07	0	0	0	0	0	0	0.13	0	0	0	0
wssdt	0.03	0	0	0	0	0	0	0	0.1	0	0	0
sndt	0.01	0	0	0	0	0	0	0	0.01	0.07	0	0
wscdt	0.01	0	0	0	0	0	0	0	0	0	0.06	0
cfhm	0.09	0.16	0	0	0	0	0	0	0	0	0	0.09

selective	O	wshm	wsnm	cfnm	wsac	wslo	cfac	wsdt	wssdt	wsndt	wscdt	cfhm
O	94.88	0.34	0.11	0.06	0.04	0.05	0.05	0.03	0.02	0	0.01	0.03
wshm	0.33	0.9	0	0	0	0	0	0	0	0	0	0.11
wsnm	0.34	0	0.64	0	0	0	0	0	0	0	0	0
cfnm	0.08	0	0.01	0.21	0	0	0	0	0	0	0	0
wsac	0.08	0	0	0	0.22	0	0.03	0	0	0	0	0
wslo	0.15	0	0	0	0	0.2	0	0	0	0	0	0
cfac	0.06	0	0	0	0.03	0	0.13	0	0	0	0	0
wsdt	0.07	0	0	0	0	0	0	0.13	0	0	0	0
wssdt	0.03	0	0	0	0	0	0	0	0.1	0	0	0
wsndt	0.01	0	0	0	0	0	0	0	0.01	0.07	0	0
wscdt	0.01	0	0	0	0	0	0	0	0	0.01	0.06	0
cfhm	0.09	0.18	0	0	0	0	0	0	0	0	0	0.07

random	O	wshm	wsnm	cfnm	wsac	wslo	cfac	wsdt	wssdt	wsndt	wscdt	cfhm
O	94.82	0.37	0.14	0.07	0.04	0.04	0.05	0.04	0.02	0.01	0.01	0.03
wshm	0.35	0.86	0	0	0	0	0	0	0	0	0	0.14
wsnm	0.34	0	0.64	0	0	0	0	0	0	0	0	0
cfnm	0.09	0	0.01	0.2	0	0	0	0	0	0	0	0
wsac	0.1	0	0	0	0.19	0	0.04	0	0	0	0	0
wslo	0.16	0	0	0	0	0.19	0	0	0	0	0	0
cfac	0.05	0	0	0	0.03	0	0.15	0	0	0	0	0
wsdt	0.07	0	0	0	0	0	0	0.13	0	0	0	0
wssdt	0.03	0	0	0	0	0	0	0	0.1	0	0	0
sndt	0.01	0	0	0	0	0	0	0	0.01	0.07	0	0
wscdt	0.01	0	0	0	0	0	0	0	0	0	0.06	0
cfhm	0.09	0.16	0	0	0	0	0	0	0	0	0	0.09

selective	O	wshm	wsnm	cfnm	wsac	wslo	cfac	wsdt	wssdt	wsndt	wscdt	cfhm
O	94.88	0.34	0.11	0.06	0.04	0.05	0.05	0.03	0.02	0	0.01	0.03
wshm	0.33	0.9	0	0	0	0	0	0	0	0	0	0.11
wsnm	0.34	0	0.64	0	0	0	0	0	0	0	0	0
cfnm	0.08	0	0.01	0.21	0	0	0	0	0	0	0	0
wsac	0.08	0	0	0	0.22	0	0.03	0	0	0	0	0
wslo	0.15	0	0	0	0	0.2	0	0	0	0	0	0
cfac	0.06	0	0	0	0.03	0	0.13	0	0	0	0	0
wsdt	0.07	0	0	0	0	0	0	0.13	0	0	0	0
wssdt	0.03	0	0	0	0	0	0	0	0.1	0	0	0
wsndt	0.01	0	0	0	0	0	0	0	0.01	0.07	0	0
wscdt	0.01	0	0	0	0	0	0	0	0	0.01	0.06	0
cfhm	0.09	0.18	0	0	0	0	0	0	0	0	0	0.07

random	O	wshm	wsnm	cfnm	wsac	wslo	cfac	wsdt	wssdt	wsndt	wscdt	cfhm
O	94.82	0.37	0.14	0.07	0.04	0.04	0.05	0.04	0.02	0.01	0.01	0.03
wshm	0.35	0.86	0	0	0	0	0	0	0	0	0	0.14
wsnm	0.34	0	0.64	0	0	0	0	0	0	0	0	0
cfnm	0.09	0	0.01	0.2	0	0	0	0	0	0	0	0
wsac	0.1	0	0	0	0.19	0	0.04	0	0	0	0	0
wslo	0.16	0	0	0	0	0.19	0	0	0	0	0	0
cfac	0.05	0	0	0	0.03	0	0.15	0	0	0	0	0
wsdt	0.07	0	0	0	0	0	0	0.13	0	0	0	0
wssdt	0.03	0	0	0	0	0	0	0	0.1	0	0	0
sndt	0.01	0	0	0	0	0	0	0	0.01	0.07	0	0
wscdt	0.01	0	0	0	0	0	0	0	0	0	0.06	0
cfhm	0.09	0.16	0	0	0	0	0	0	0	0	0	0.09

selective	O	wshm	wsnm	cfnm	wsac	wslo	cfac	wsdt	wssdt	wsndt	wscdt	cfhm
O	94.88	0.34	0.11	0.06	0.04	0.05	0.05	0.03	0.02	0	0.01	0.03
wshm	0.33	0.9	0	0	0	0	0	0	0	0	0	0.11
wsnm	0.34	0	0.64	0	0	0	0	0	0	0	0	0
cfnm	0.08	0	0.01	0.21	0	0	0	0	0	0	0	0
wsac	0.08	0	0	0	0.22	0	0.03	0	0	0	0	0
wslo	0.15	0	0	0	0	0.2	0	0	0	0	0	0
cfac	0.06	0	0	0	0.03	0	0.13	0	0	0	0	0
wsdt	0.07	0	0	0	0	0	0	0.13	0	0	0	0
wssdt	0.03	0	0	0	0	0	0	0	0.1	0	0	0
wsndt	0.01	0	0	0	0	0	0	0	0.01	0.07	0	0
wscdt	0.01	0	0	0	0	0	0	0	0	0.01	0.06	0
cfhm	0.09	0.18	0	0	0	0	0	0	0	0	0	0.07

Overview

- Introduction
  - Approach & Results
- Discussion
  - Alternative Selection Metrics
  - Costing Active Learning
  - Error Analysis
- **Conclusions**

Conclusions

AL for IE with a Committee of Classifiers:

- Approach using KL-divergence to measure disagreement amongst MEMM classifiers
  - Classification framework: simplification of IE task
- Ave. Improvement: 1.3 absolute, 2.1 % f-score

Suggestions:

- Interaction between AL methods and text-based cost estimates
  - Comparison of methods will benefit from real cost information…
- Full simulation?

Thank you

The SEER/EASIE Project Team

Edinburgh:

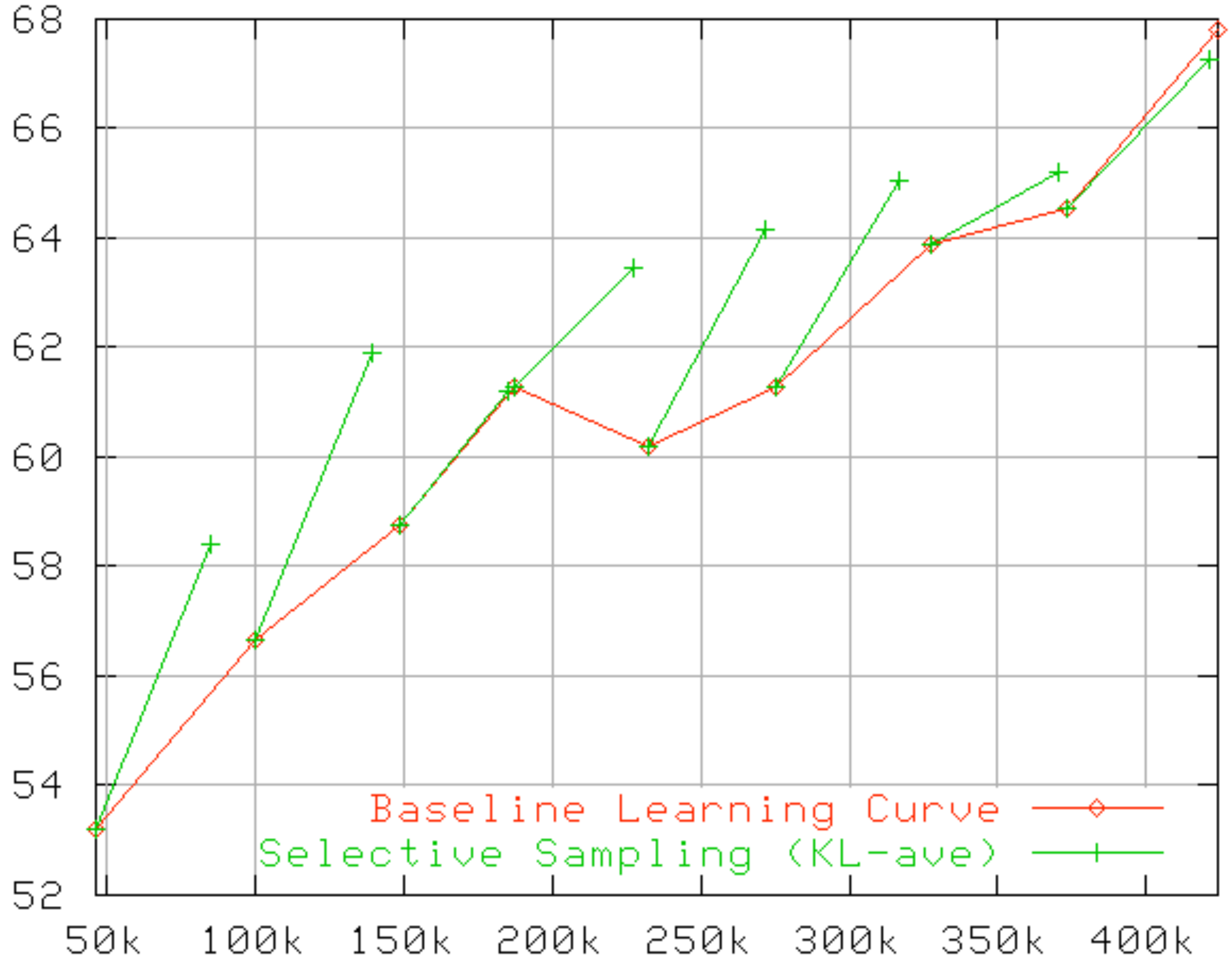
Bea Alex, Markus Becker, Shipra Dingare, Rachel Dowsett, Claire Grover, Ben Hachey, Olivia Johnson, Ewan Klein, Yuval Krymolowski, Jochen Leidner, Bob Mann, Malvina Nissim, Bonnie Webber

Stanford:

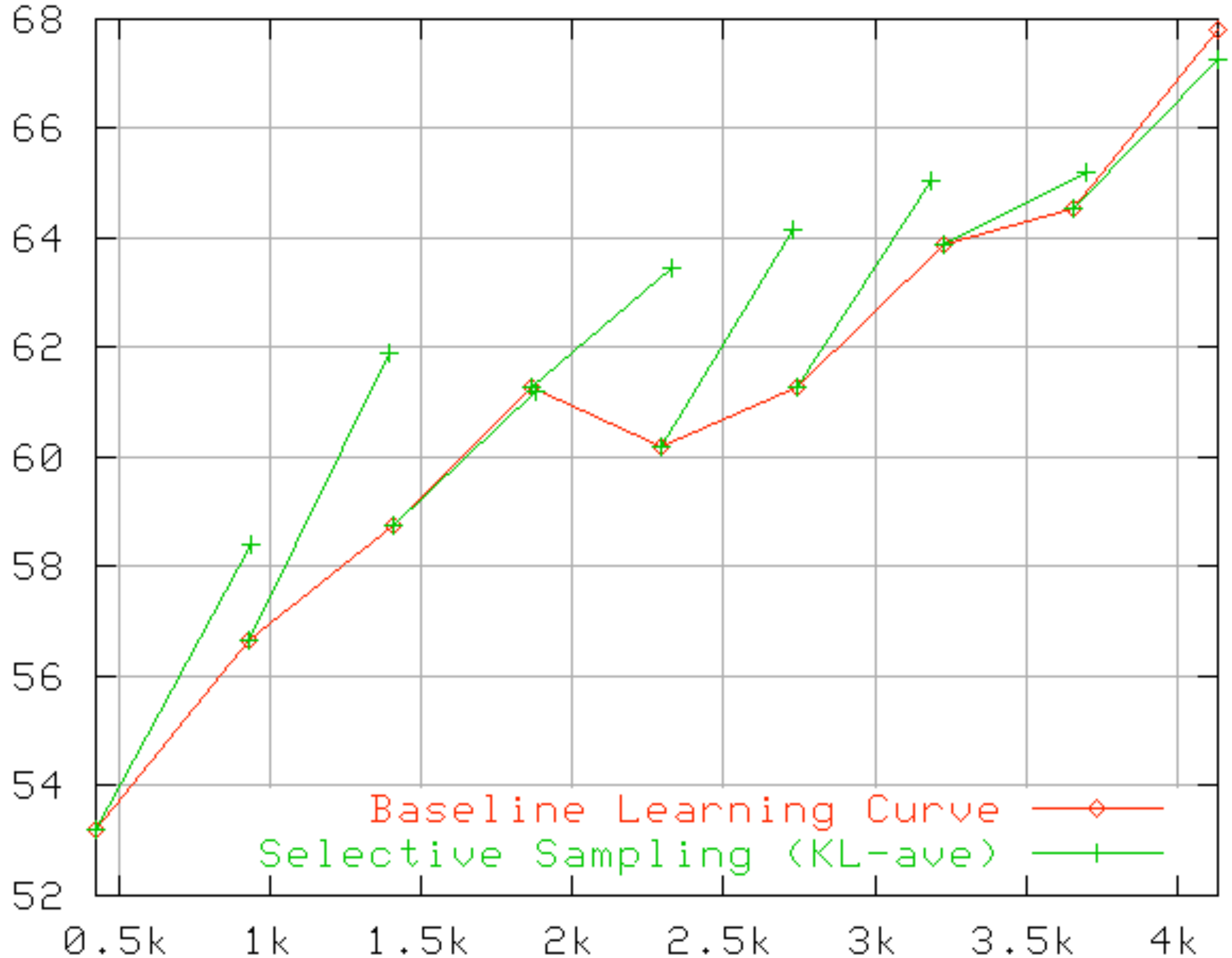
Chris Cox, Jenny Finkel, Chris Manning, Huy Nguyen, Jamie Nicolson

More Results

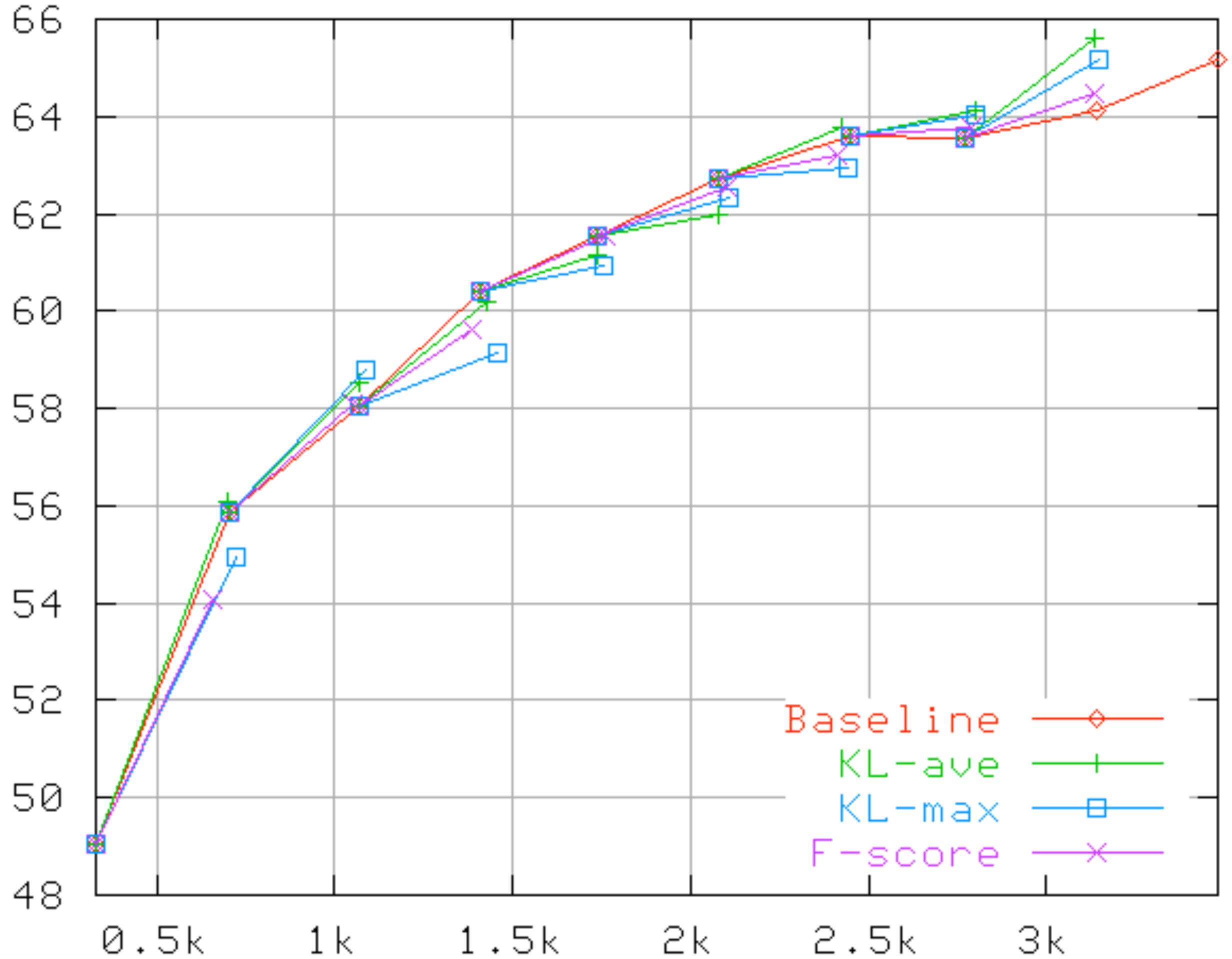
Evaluation Results: Tokens



Evaluation Results: Entities



Entity Cost Metric (Dev)



More Analysis

Boundaries: Acc+class/Acc-class

Round	Random	Selective
1	0.974/0.970	0.975/0.970
4	0.977/0.971	0.977/0.972
8	0.978/0.973	0.979/0.975

Boundaries: Full/Left/Right F-score

Round	Random	Selective	Δ
1	0.564/0.593/0.588	0.568/0.594/0.593	0.004/0.001/0.018
4	0.623/0.648/0.647	0.619/0.643/0.643	-.004/-.005/-.004
8	0.648/0.669/0.676	0.663/0.684/0.690	0.015/0.015/0.013