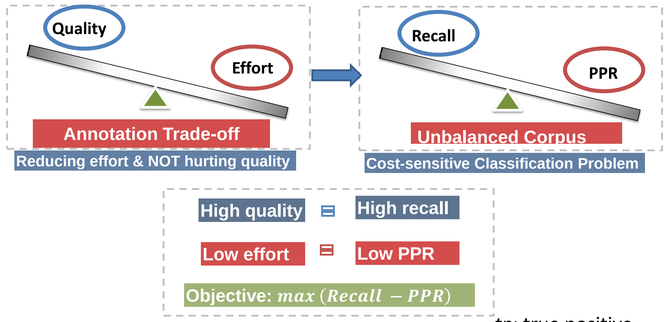
Reducing Annotation Effort on Unbalanced Corpus based on Cost Matrix*

Wencan Luo[1], Diane Litman[1,2] and Joel Chan[2]
[1]Department of Computer Science, University of Pittsburgh
[2]Learning Research and Development Center, University of Pittsburgh
{wencan,litman}@cs.pitt.edu, {joc59}@pitt.edu

Motivation

- High-quality annotated corpora are desirable.
- Annotation is *tedious* and *costly.*
- It's a tradeoff between *data quality* and *human effort.*
- Reducing annotation effort but not hurting the quality?



$PPR = \frac{tp+fp}{N}$ $Recall = \frac{tp}{tp+fn}$

tp: true positive
fp: false positive
N: total number

Skewed Distribution (unbalanced degree: 3% ~ 24%)

- *text classification (Forman, 2003)*
- *information extraction (Hoffmann et al., 2011)*
- *emotion detection (Ang et al., 2002; Alm et al., 2005)*
- *sentiment classification (Li et al., 2012)*
- *polarity of opinion (Carvalho et al., 2011)*
- *uncertainty and correctness of student answers in tutoring dialogue systems (Forbes-Riley and Litman, 2011; Dzikovska et al., 2012)*

Related Work

- Semi-supervised learning methods
  1. *active learning* (Cohn et al., 1994; Zhu and Hovy, 2007; Zhu et al., 2010)
  2. *co-training* (Blum and Mitchell, 1998)
  3. *self-training* (Mihalcea, 2004)
- Supervised methods + manual checking

  pre-annotation (Brants and Plaehn, 2000; Chiou et al., 2001; Xue et al., 2002; Ganchev et al., 2007; Chou et al., 2006; Rehbein et al., 2012)

Our Approach

- Belongs to pre-annotation
- Annotation Steps:
  1. build a **high-recall classifier**
  2. apply to the rest of the unlabeled data
  3. manually check every **positive** label

Cost Matrix **High-Recall Classifier**

- Cost-sensitive problem
  - a trivial solution : classify all instances as '1'
  - 100% recall, but high positive predict → no cost savings
- Objective: high recall & low positive predict rate (PPR) $PPR = \frac{tp+fp}{N}$

	Actual class 1	Actual class 0
Predicted class 1	C_{tp}	C_{fp}
Predicted class 0	C_{fn}	C_{tn}

$$Recall = \frac{tp}{tp+fn}$$

Intrinsic Evaluation

- Transcribed human-human dialogues
- Annotated by two coders (kappa = 0.75)
  - ✓ for presence/absence of uncertainty

Speaker	Utterance	Uncertainty?
S6	You can't see the forest through the trees.	No
S1	I'm not quite sure	Yes

	# of Utterances	Unbalanced Degree
Train	12,331	13.3%
Test	1,558	14.2%
Unlabeled	42,641	?

Basic Classifier

- Feature Set
  - 133 Keywords/Phases

High Recall Classifier

- $C_{fn} = 15$ (chosen based on train set)
- Up to 80% annotation effort reduction

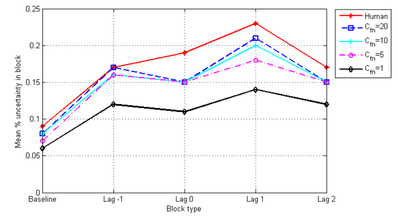
	Recall	Positive Predict Rate
Basic Classifier	**67.4%**	**10.9%**
Cost-Sensitive Classifier	**91.0%**	**20.7%**

Extrinsic Evaluation

- Concerns about the data quality
  - ✓ missed some fn
  - ✓ changed the distribution of classes
- Extrinsic task
  - ✓ Replicate the analysis of uncertainty level change with the use of analogies (Chan et al., 2012)

Results

- Under two conditions
  - manual label (**black line**)
  - predicted label and manually check ($C_{fn} = 5,10,20$)
- **NOT** substantially alter or miss known statistical effects



Conclusion

- An annotation scheme based on cost matrix
  - ✓ lowers the threshold to build a high-recall classifier
  - ✓ reduces significant annotation effort (by checking only the positive predictions)
  - ✓ does not sacrifice data quality
- Future work
  - ✓ experiment with different tasks
  - ✓ extend to multi-class classification tasks
  - ✓ explore effects of the degree of unbalance

*** This work is published in NAACL-SRW, 2013**
Wencan Luo, Diane Litman and Joel Chan. Reducing Annotation Effort on Unbalanced Corpus based on Cost Matrix. ***Proceedings of the NAACL HLT Student Research Workshop*****, Atlanta, GA, June.**