Active Learning Strategies Based on Text Informativeness

Ruide Li (Kyoto University)

Yoko Yamakata (The University of Tokyo)

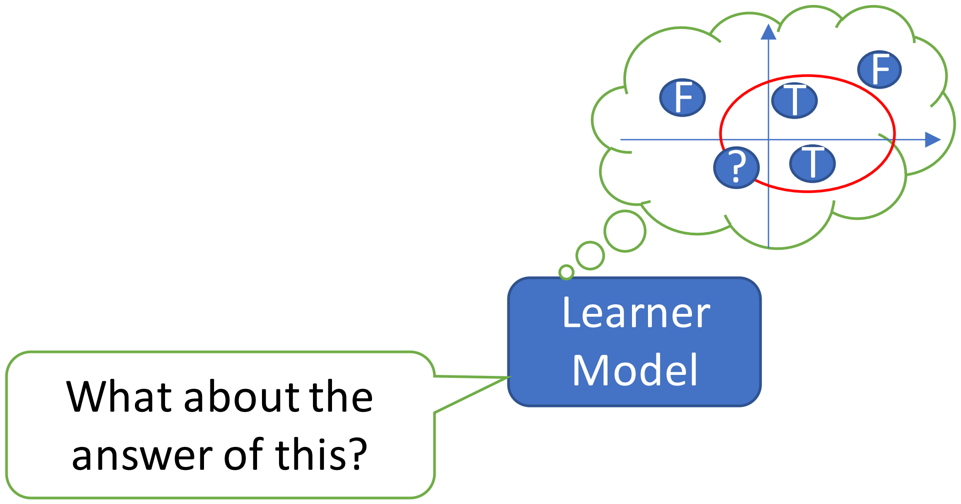
Keishi Tajima (Kyoto University)

What Is Active Learning

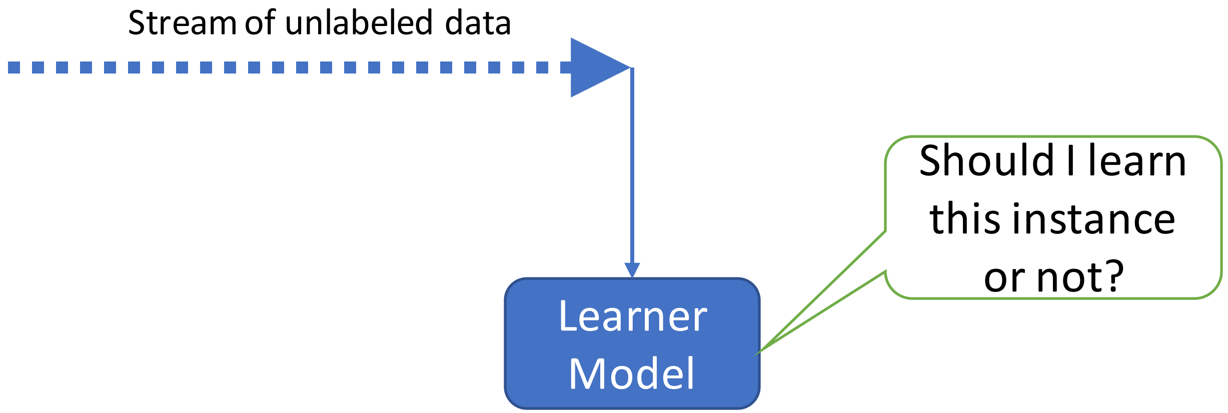
- Problem in supervised machine learning:
  - Unlabeled data is abundant, while annotation cost is high
- What if a model can ask its ”supervisor” for labels?
  - Actively choose data for labeling to learn

Pool-Based Active Learning

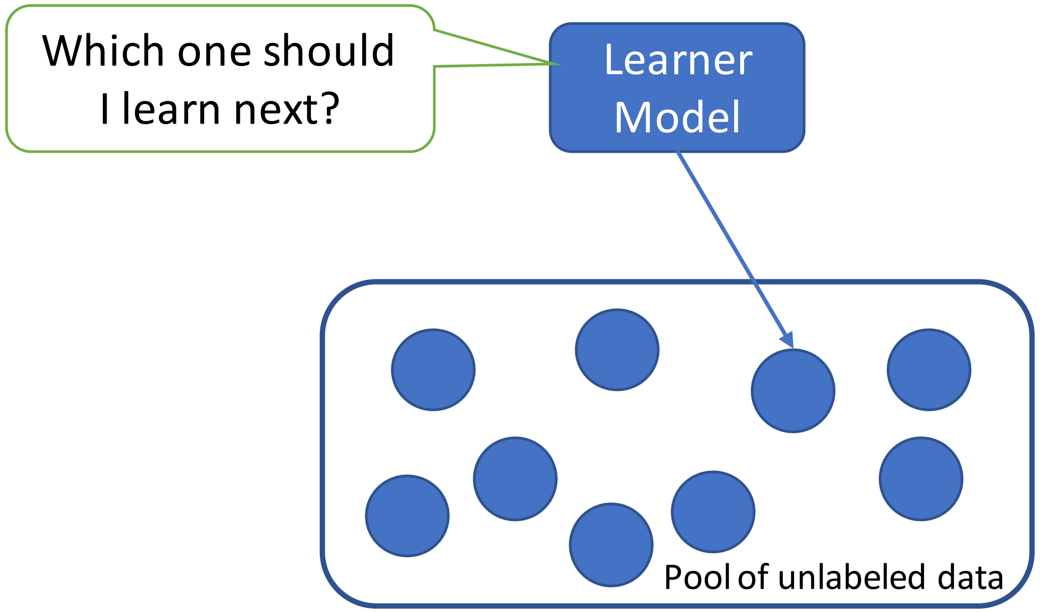
- 3 main types of Active Learning:
  - Membership Query Synthesis
  - Pool-Based Samplling
  - Stream-Based Selective Sampling



Membership Query Synthesis



Stream-based



Pool-based

What If Specific Data Domain Is Given

- Given a fixed pool of text data, is there any approach which the learner can take advantage of?
  - Fixed pool: pool-based Active Learning
  - Text data: language model features

Introduction

Problem Settings

Related Work

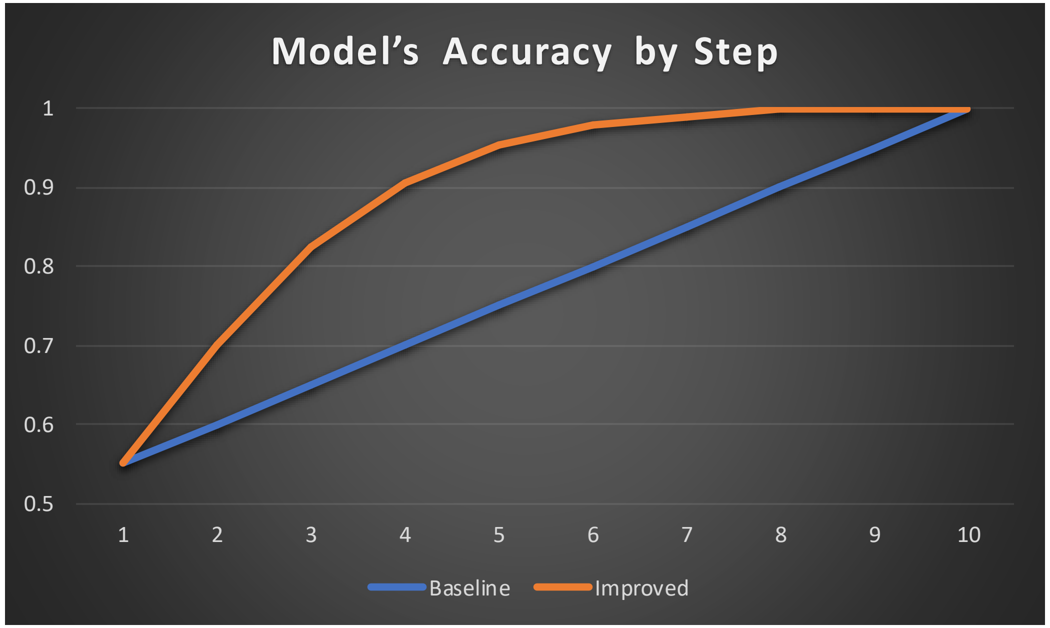
Proposed Methods

Experiments

Conclusion

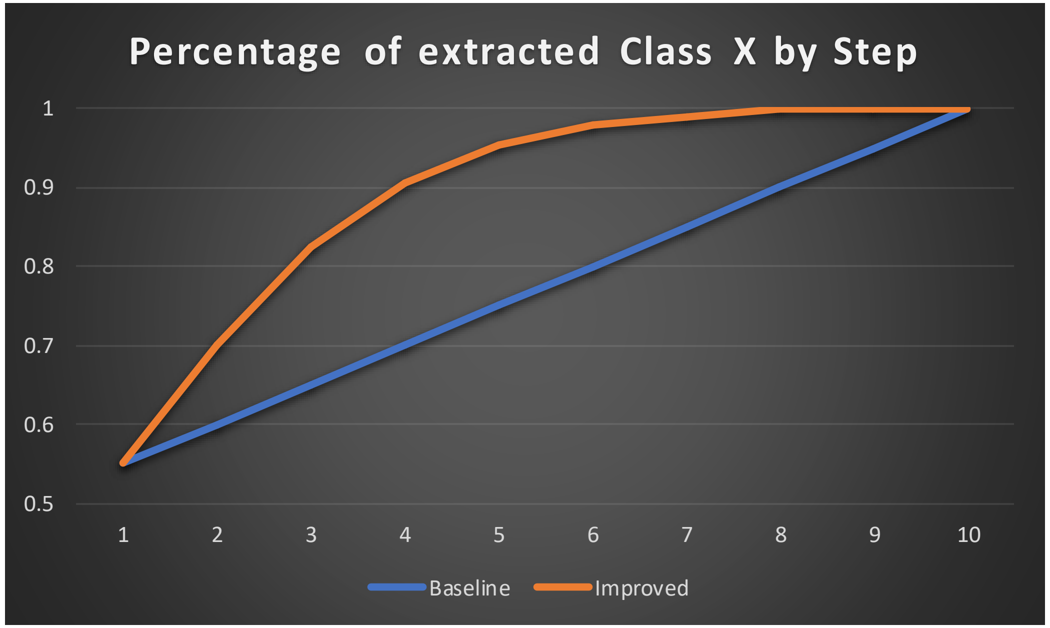
Standard Active Learning

- Improve the model's accuracy with as few human annotation as possible
  - Desired output: trained model



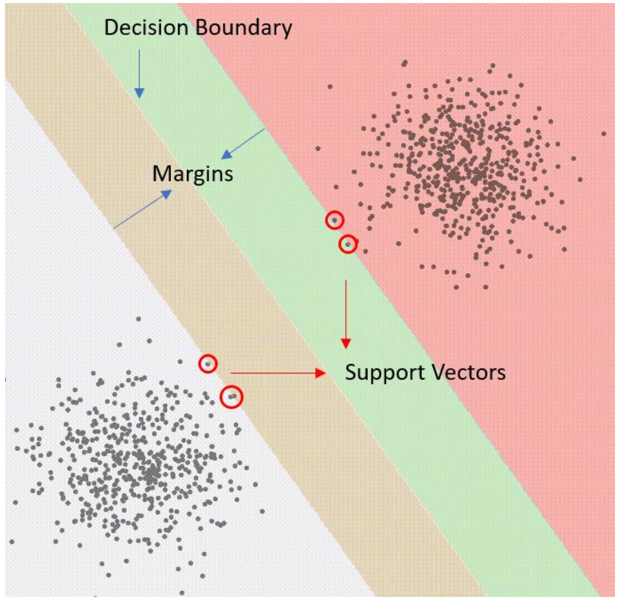
Learn-to-Enumerate

- Extract a certain class of data from the unlabeled data pool with as few human annotation as possible
  - Desired output: all data of a specific class



Standard Active Learning

- Uncertainty Sampling
  - label those items for which the current model is least certain as to what the ground truth should be
  - In SVM, it is tantamount to search for the support vectors ASAP



D. D. Lewis and W. A. Gale, “A sequential algorithm for training text classifiers,” in *Proc. of SIGIR*, 1994, pp. 3–12

Learn-to-Enumerate

- ε-greedy exploitation and exploration
  - With probability ε, do exploration, i.e., the current model is least confident
  - With probability 1 - ε, do exploitation, i.e., the current model is most confident
- Exploitation-only strategy gives the best result

P. Jörger, Y. Baba, and H. Kashima, “Learning to enumerate,” in *Proc. of Intl. Conf. on Artificial Neural Networks, Part I*, 2016, pp. 453–460.

Introduction

Problem Settings

Related Work

Proposed Methods

Experiments

Conclusion

Query Strategy Design of Text Data

- Manage the unlabeled data in an certain order to achieve our goal
- Deside the definition of informativeness (primitive methods)
  - Unique word count
  - Sum of TF-IDF
  - Sum of TF-IDF of unseen words
  - Norm of Doc2Vec
- Combine our primitive methods with a baseline method in each problem setting

Unique Word Count

- Count unique words in each document
  - Long articles with many different words are difficult to understand
  - If the document has many non-repetitive words, the document is informative

Sum of TF-IDF

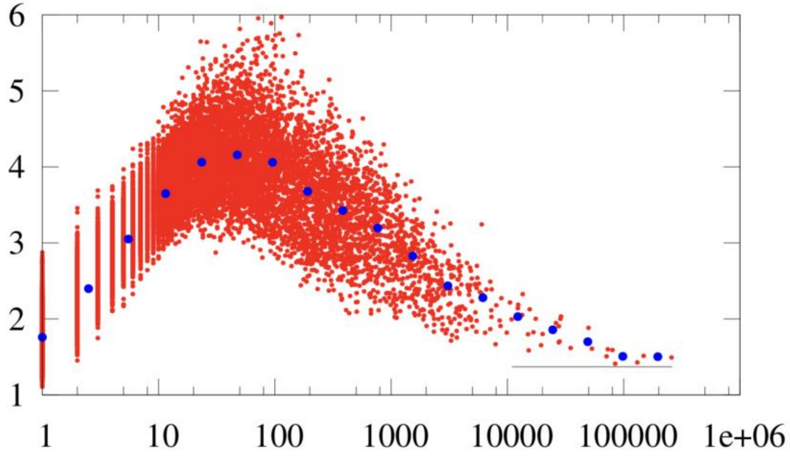
- Term frequency–inverse document frequency
  - Term frequency: $\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$
  - Inverse document frequency: $\mathrm{idf}(\mathrm{t},\mathrm{D}) = \log \frac{N}{|\{d \in D : t \in d\}|}$
  - TF-IDF: $\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$
- Sum up TF-IDF scores of all words in a document
- However, this calculation is too much affected by very unusual words (very large IDF)
  - Only use top-*k* TF-IDF scores

Sum of TF-IDF of Unseen Words

- If some words are already learnt, it is not necessary to learn these words repetitively
- Only calculate TF-IDF scores of unprecedented words

Norm of Embedding Vector (Word2Vec)

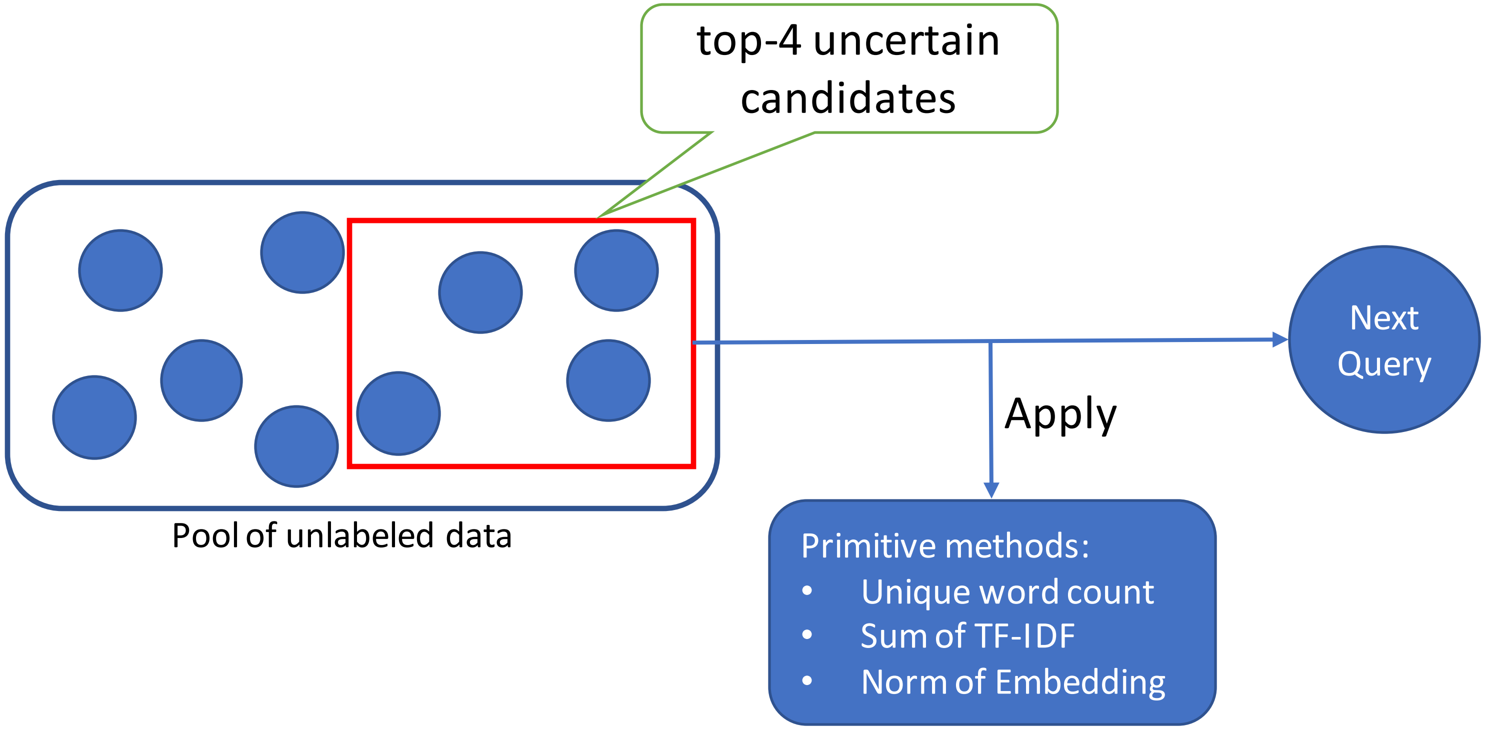
- When TF is less than a certain threshold, norm of word embedding increases as TF rises
  - The word vector is updated frequently during training
- When TF rises further, the norm will decrease
  - The word vector is updated so frequently that is stretched flat
  - Extremely frequent words fit many context
- Extend this attribute to document, using Doc2Vec



Word embedding norm by TF

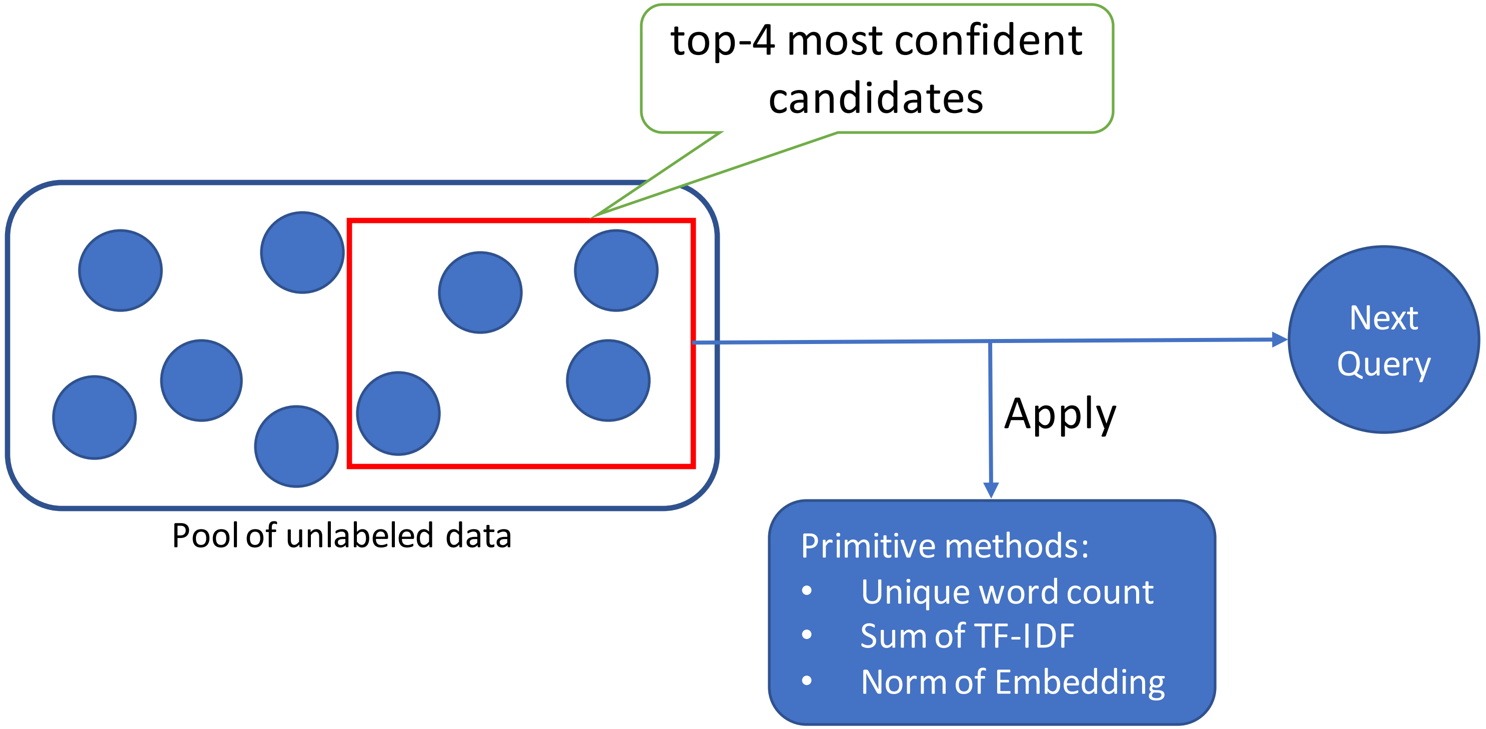
Combined with Uncertainty Sampling

- In uncertainty sampling, instead of calculate the most uncertain item, we make it yield top-*k* candidates
- Apply primitive approaches on these candidates



Combined with Exploitation-Only

- In exploitation-only ε-greedy strategy, instead of calculate the most confident item, we make it yield top-*k* candidates
- Apply primitive approaches on these candidates



Introduction

Problem Settings

Related Work

Proposed Methods

Experiments

Conclusion

Experiment Detail

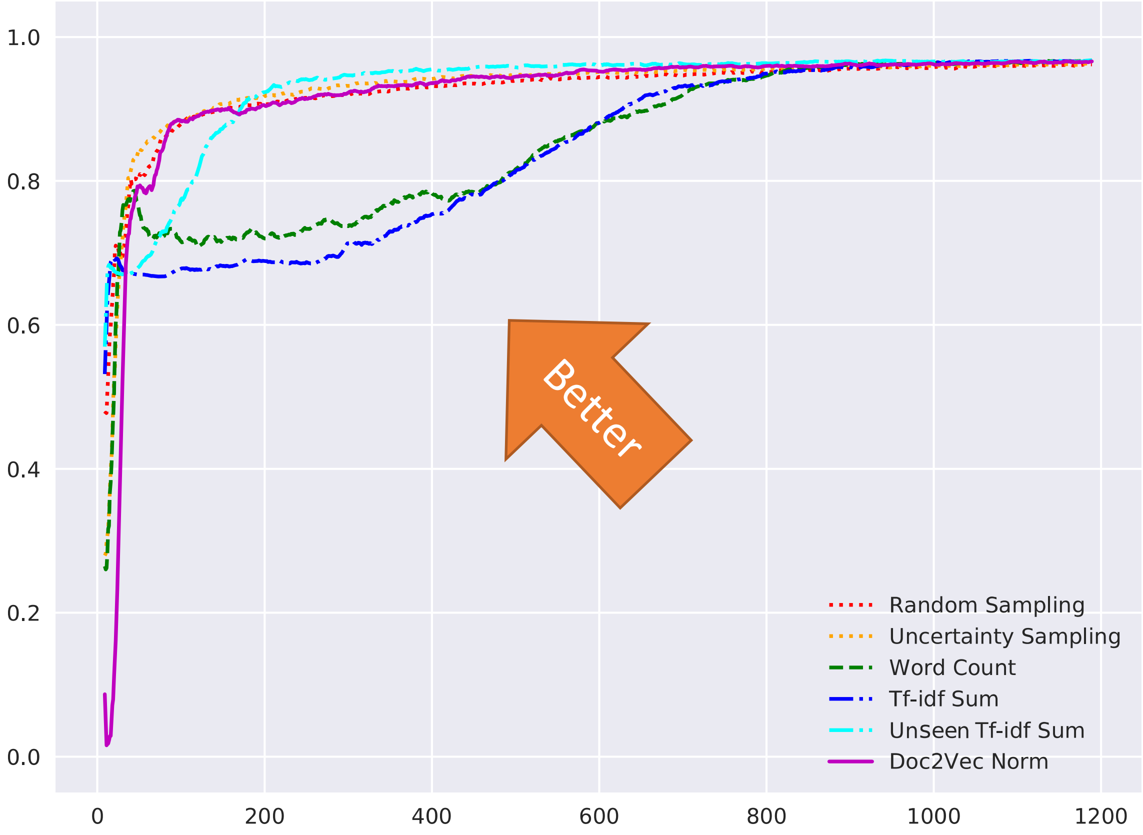
- When selecting top-k words having the highest TF-IDF values in our method, we selected 20 words
- In the combination methods, we first choose top 10 candidates
- Learner model, Support Vector Machine (SVM) with default hyper-parameters in SciKit-Learn
  - Computational cost
  - Small dataset size
- Baseline
  - Standard Active Learning: uncertainty sample
  - Learn-to-Enumerate: exploitation-only ε-greedy strategy

Description of Dataset 1

- SMS Spam Collection Dataset
  - UCI Machine Learning Repository
- Spam: 50%, ham: 50%
- Learn-to-enumerate target: spam

Results on Dataset 1: Primitive Methods

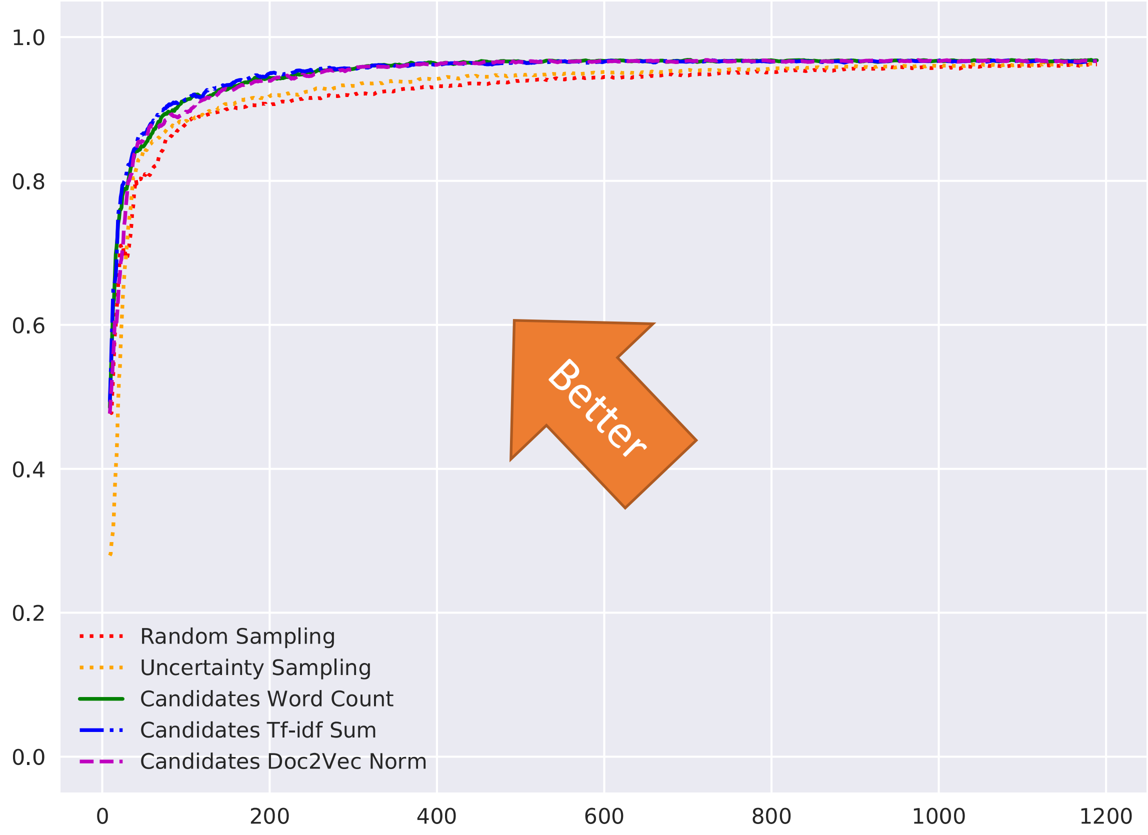
- Primitive methods showed worse result than baseline



F1 score by step

Results on Dataset 1: Combined Methods

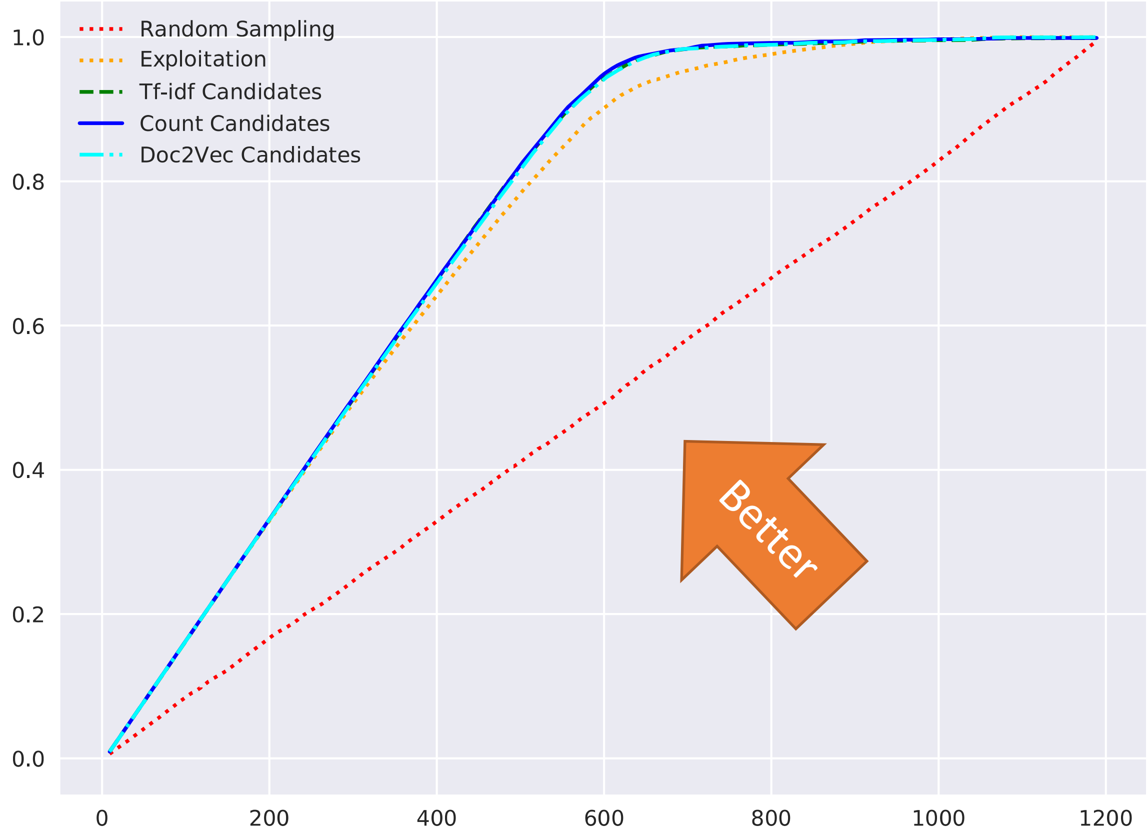
- Our methods consistantly outperformed baseline



F1 score by step

Results on Dataset 1: Learn-to-Enumerate

- Our methods consistantly outperformed baseline



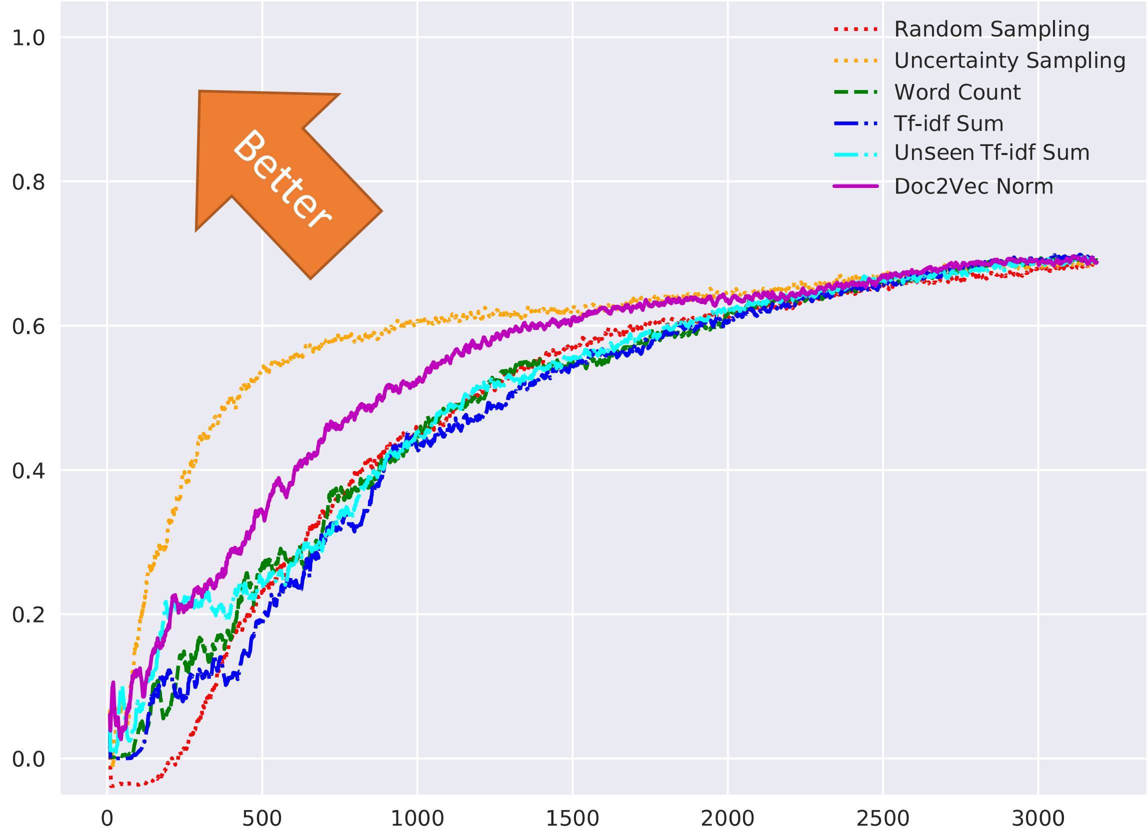
Proportion of target class extracted by step

Description of Dataset 2

- Binary sentiment classification of movie reviews
  - Large Movie Review Dataset v1.0
- Positive: 20%, negative: 80%
- Learn-to-enumerate target: positive

Results on Dataset 2: Primitive Methods

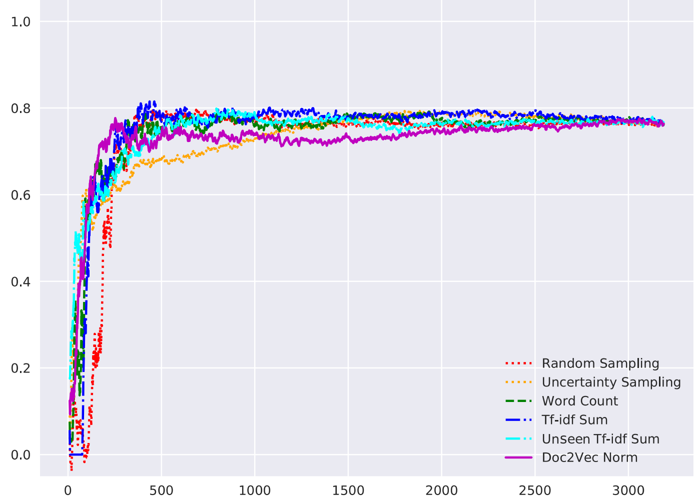
- Primitive methods showed worse result than baseline



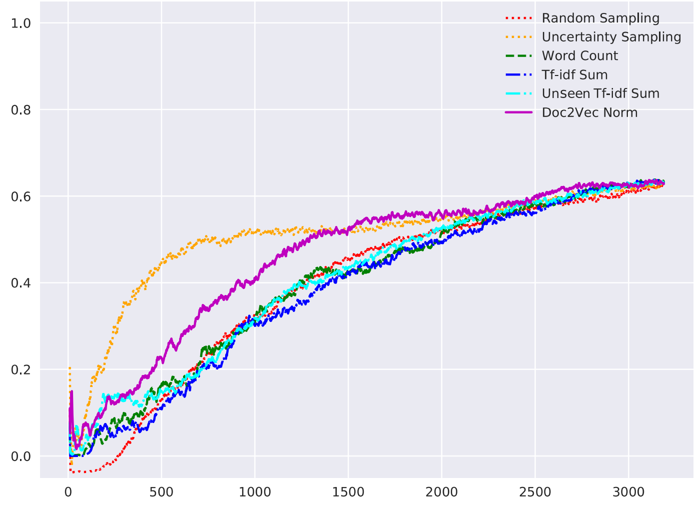
F1 score by step

Results on Dataset 2: Primitive Methods

- Our methods gave higher score on the opposite class



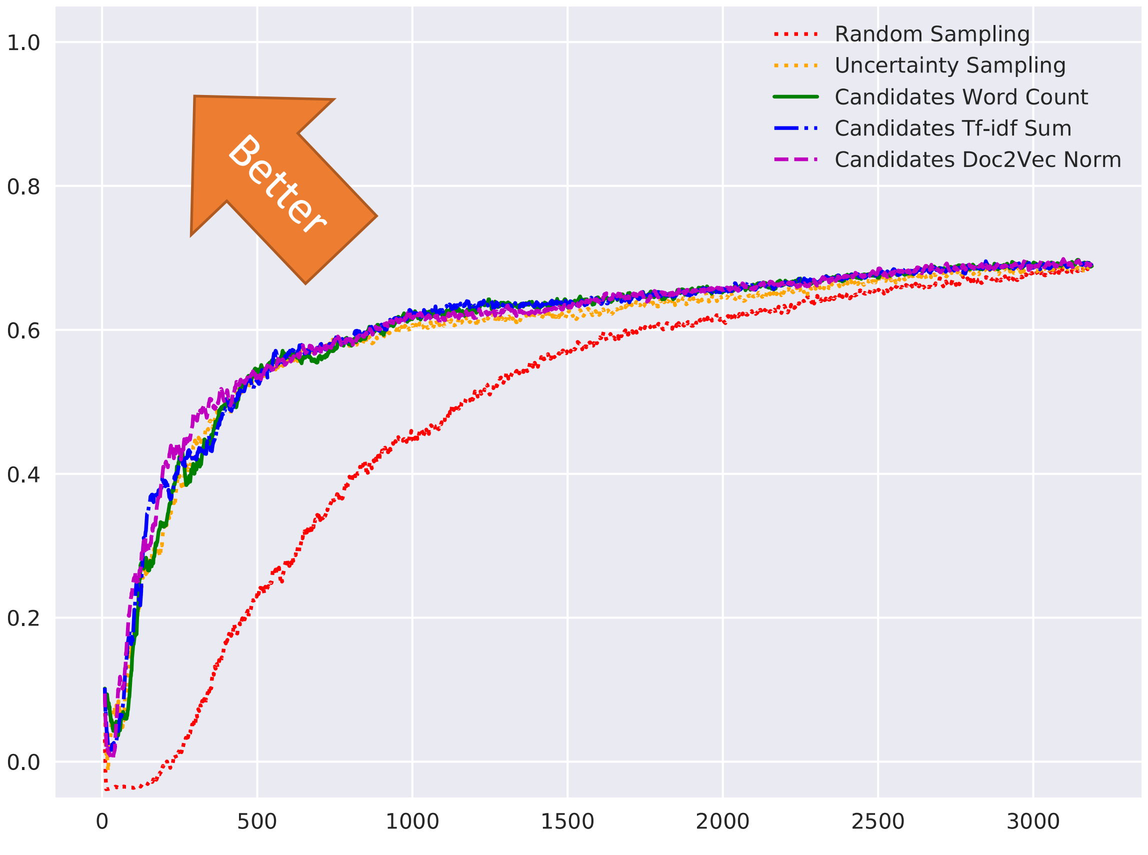
Precision by step



Recall by step

Results on Dataset 2: Combined Methods

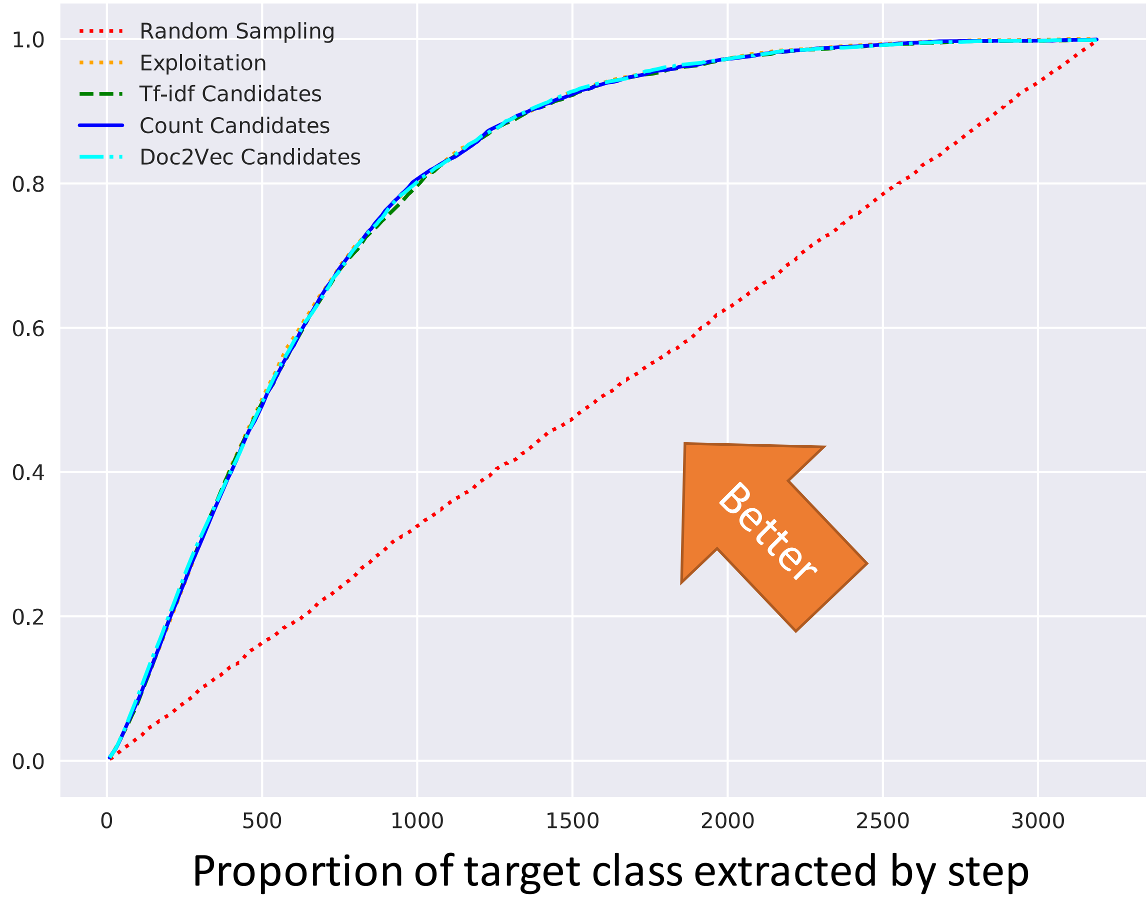
- Doc2Vec combination method outperformed baseline by a narrow but consistent margin



F1 score by step

Results on Dataset 2: Learn-to-Enumerate

- Our methods performed equally as baseline due to property of the dataset



Introduction

Problem Settings

Related Work

Proposed Methods

Experiments

Conclusion

Conclusion

- We proposed methods that utilize features specific to text data
  - Unique word count
  - Sum of TF-IDF
  - Sum of TF-IDF of unseen words
  - Norm of Doc2Vec
- Combination methods
  - Combine with uncertainty sampling to solve standard active learning problem
  - Combine with exploitation-only ε-greedy strategy to solve learn-to-enumerate problem

Standard Active Learning

- Our primitive did not always outperform uncertainty sampling
- Our combination methods outperformed it with a small but consistent margin

Learn-to-Enumerate

- Our methods outperformed the exploitation-only strategy in the experiment with Dataset 1
  - Our methods have advantage due to data property
- Our methods yielded equal result as exploitation-only strategy in the experiment with Dataset 2
  - Our methods have disadvantage due to data property
- Our methods generally have superiority over exploitation-only strategy