CS11-737 Multilingual NLP

Active Learning

Graham Neubig

Carnegie Mellon University

Language Technologies Institute

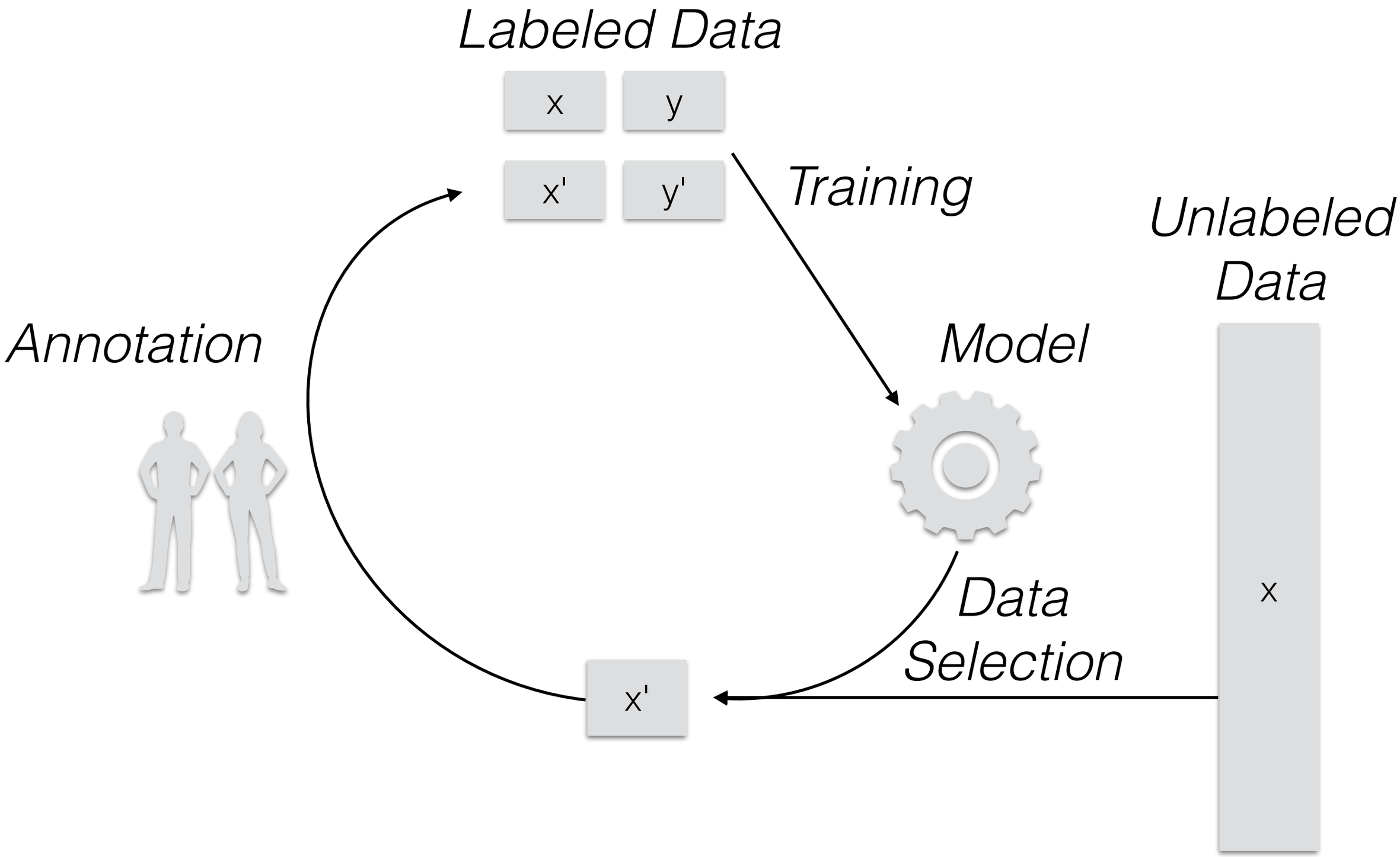
Site

http://phontron.com/class/multiling2022/

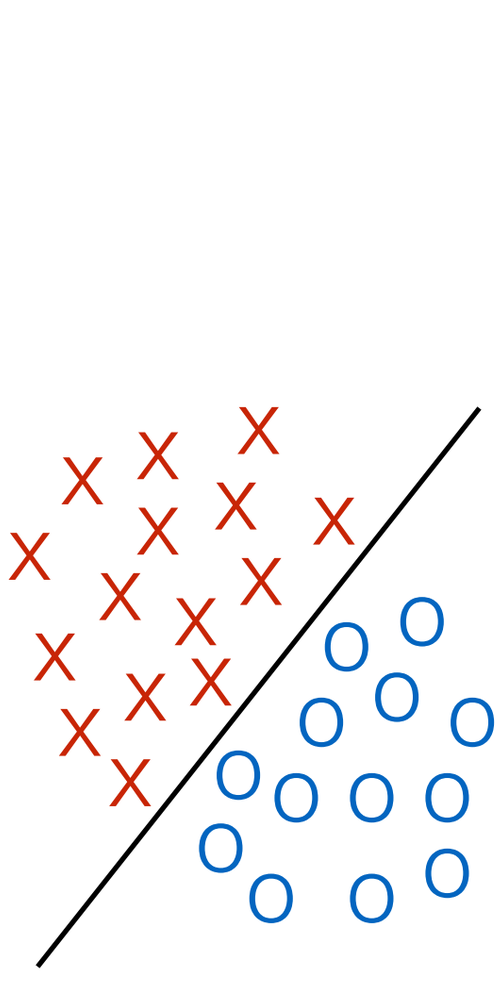
Types of Learning

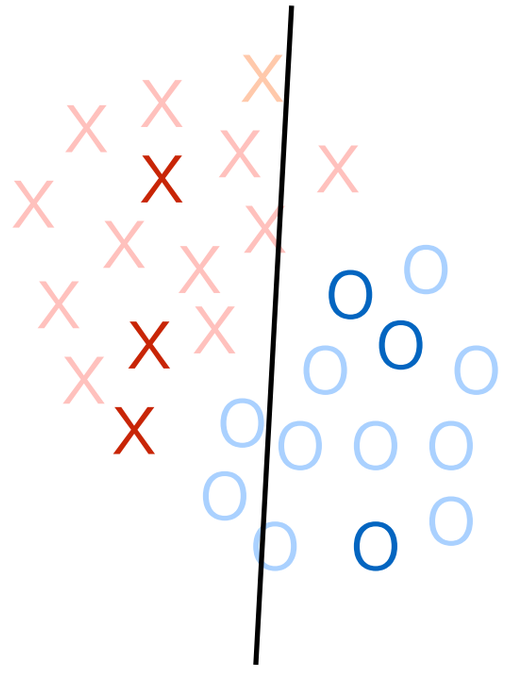
- **Supervised:** Learn from input-output pairs *<x,y>*
- **Unsupervised:** Learn from inputs only *x*
- **Semi-supervised:** Learn from both
- **Active:** query a human annotator to efficiently generate *<x,y>* examples from *x*

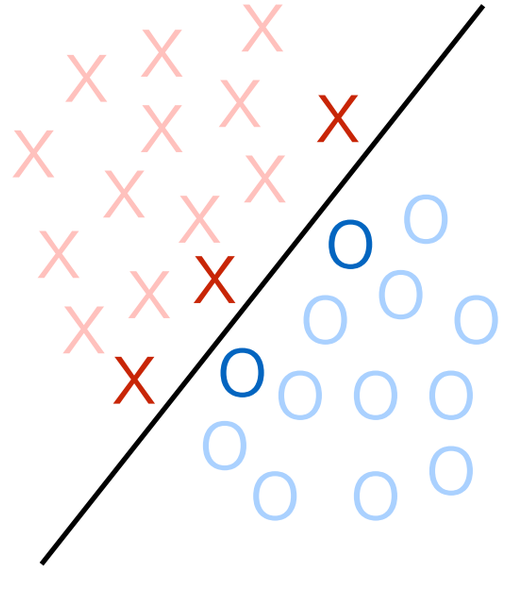
Active Learning Pipeline



Why Active Learning?







Fundamental Ideas

- **Uncertainty:** we want data that are *hard* for our current models to handle

- **Representativeness:** we want data that are *similar* to the data that we are annotating

Uncertainty/Representativeness Criteria: Example of Classification

Uncertainty Paradigms

- **Uncertainty Sampling:** Find ones where models are most uncertain

- **Query by Committee:** Use different classification models and measure agreement

Uncertainty Sampling Criteria

- **Entropy:** larger entropy = more uncertain

$$H(x) = -\sum_y P(y|x) \log P(y|x)$$

- **Top-1 confidence:** lower top-1 confidence = more uncertain

$$\hat{y} = \underset{y}{\operatorname{argmax}} \log P(y|x)$$

$$\text{top1}(x) = \log P(\hat{y}|x)$$

- **Margin:** smaller difference between first and second candidates = more uncertain

$$\text{margin}(x) = \log P(\hat{y}|x) - \max_{y \neq \hat{y}} \log P(y|x)$$

Tong, Simon, and Daphne Koller. "Support vector machine active learning with applications to text classification." *Journal of machine learning research* 2.Nov (2001): 45-66.

Culotta, Aron, and Andrew McCallum. "Reducing labeling effort for structured prediction tasks." *AAAI*. Vol. 5. 2005.

Query by Committee

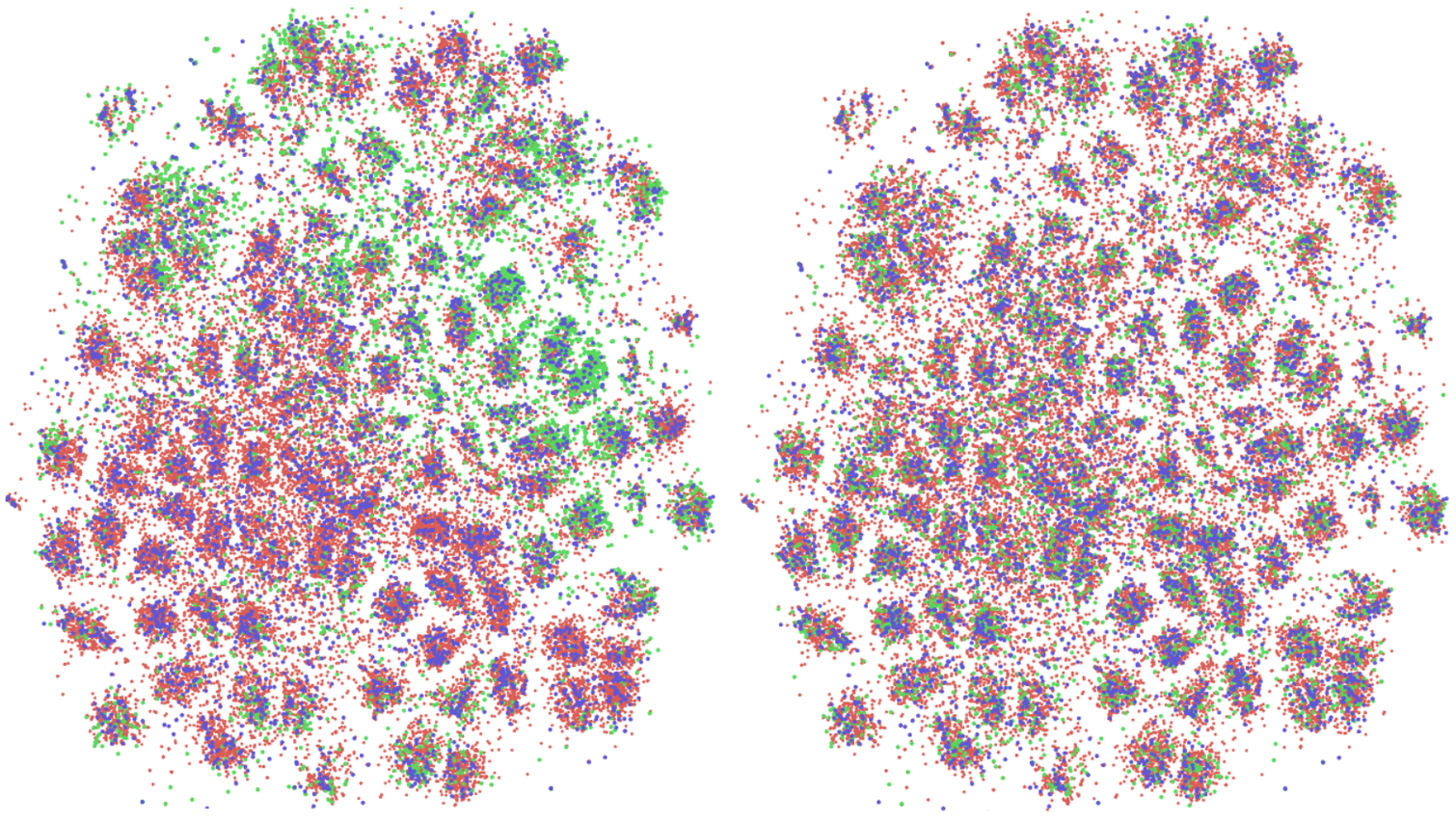
- Run multiple models and measure the disagreement
- Can be combined with standard ensembling methods (e.g. bagging and boosting)

Seung, H. Sebastian, Manfred Opper, and Haim Sompolinsky. "Query by committee." *Proceedings of the fifth annual workshop on Computational learning theory*. 1992.

Representativeness

- How can we classify examples as being "similar to many others"?

- In simple feature vectors: high overlap in vector space



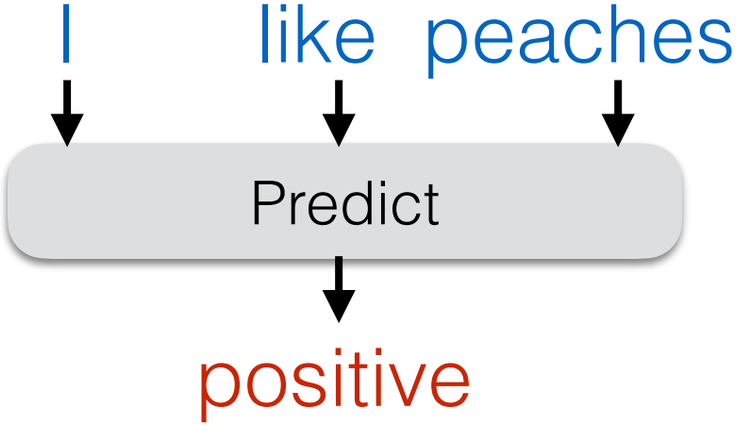
(a) Uncertainty Oracle (b) Our Method

Sener, Ozan, and Silvio Savarese. "Active learning for convolutional neural networks: A core-set approach." *arXiv preprint arXiv:1708.00489* (2017).

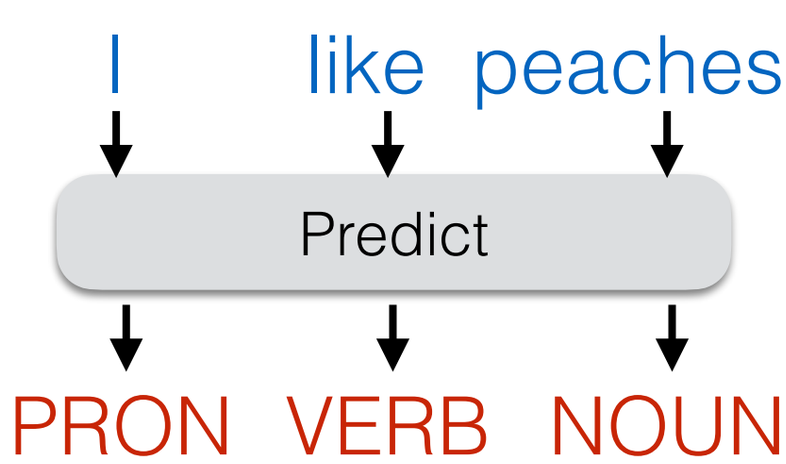
Active Learning Strategies for Text

Prediction Paradigms for Language Tasks

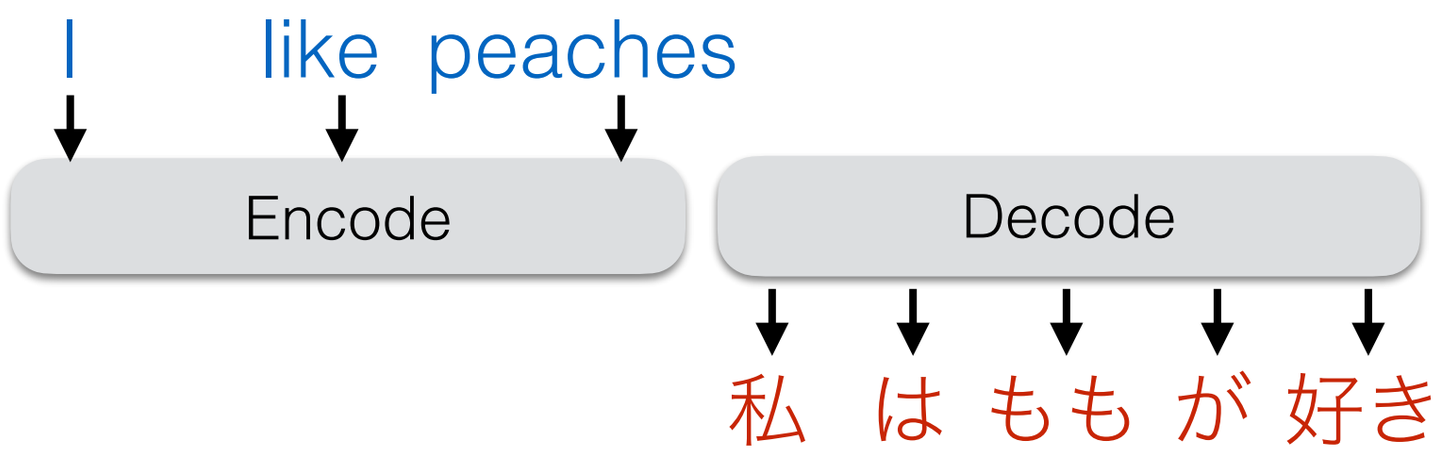
- **Text classification**



- **Tagging**

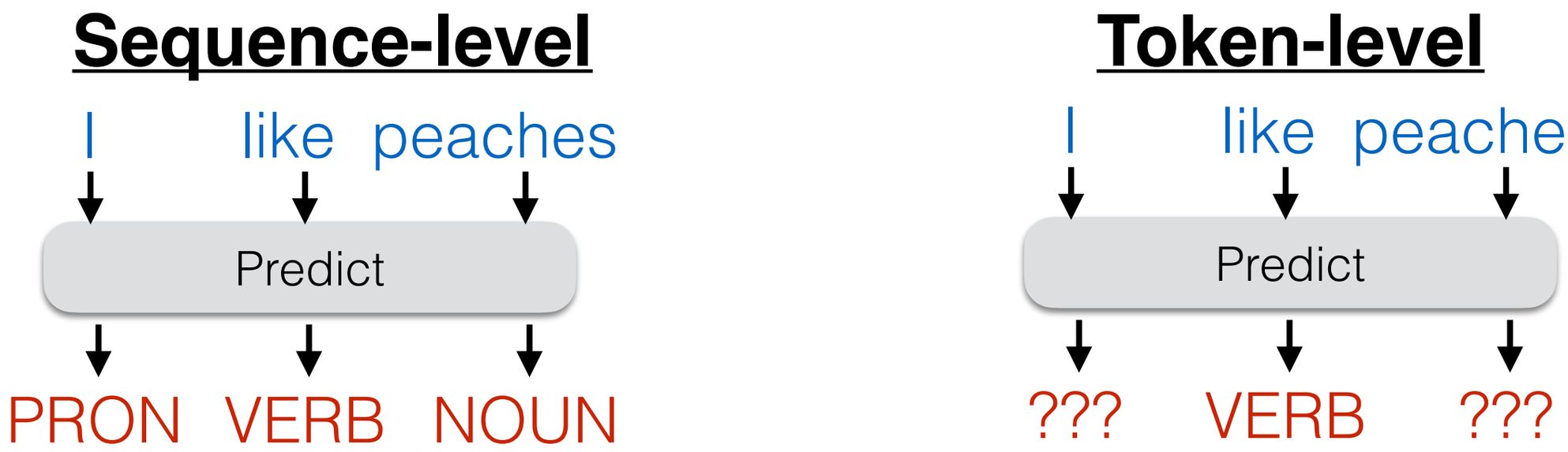


- **Sequence-to-sequence**



Sequence / Token Level

- For sequence labeling and sequence-to-sequence, can do annotation at different levels



- Token level makes it possible to annotate just difficult parts of sentences -> time savings?

- But it requires strategies to learn from individual examples

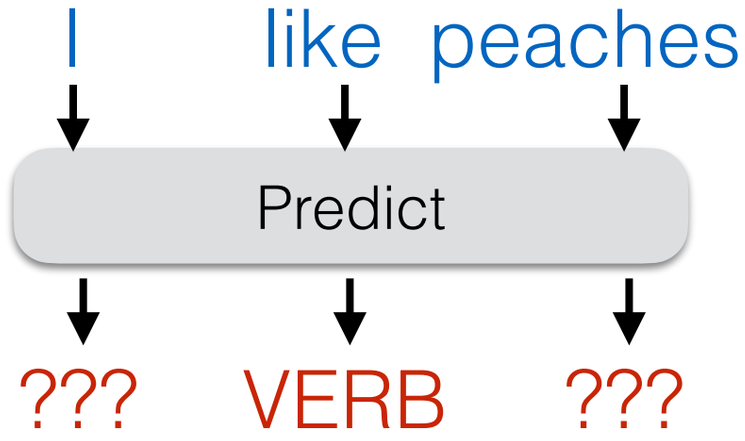
Sequence-level Uncertainty Measures

- **Top-1 confidence:** Trivial to apply
- **Margin:** Find 1-best and 2-best, margin
- **Sequence-level entropy:** Non-trivial to enumerate, can be enumerated over *n*-best candidates

Settles, Burr, and Mark Craven. "An analysis of active learning strategies for sequence labeling tasks." *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. 2008.

Training on Token Level

- How can we train on partial data?



- **Unstructured predictors:** Treat each prediction as independent, and train on only annotated labels

- **Marginalization:** For structured prediction methods (e.g. CRFs) we can sum over all unlabeled tokens

Tsuboi, Yuta, et al. "Training conditional random fields using incomplete annotations." *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. 2008.

Neubig, Graham, Yosuke Nakata, and Shinsuke Mori. "Pointwise prediction for robust, adaptable Japanese morphological analysis." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 2011.

Token-level Representativeness Metrics

- Can accumulate uncertainty across token instances
  (e.g. uncertainty of "run" will be sum of uncertainties across all instances)

- Select a representative instance of that token

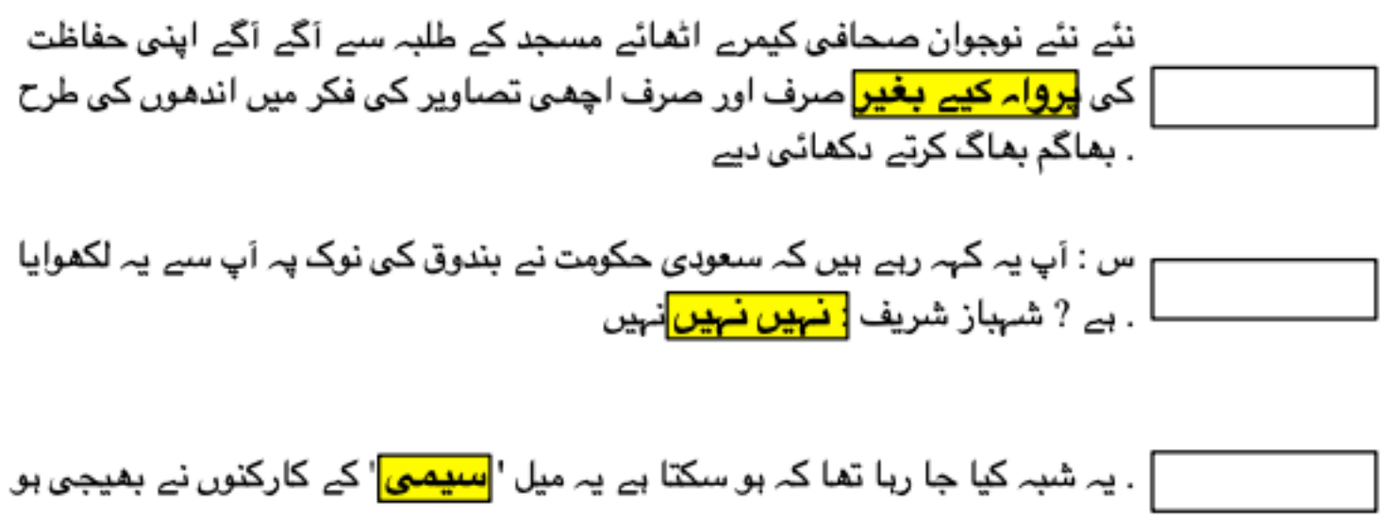
Fang, Meng, and Trevor Cohn. "Model transfer for tagging low-resource languages using a bilingual dictionary." *arXiv preprint arXiv:1705.00424* (2017).

Sequence-to-sequence Uncertainty Metrics

- **Sequence-level:** e.g. back-translation likelihood

$$\mathrm{BTL}(x) = \log P(x|\hat{y})$$

- **Phrase level:** e.g. most frequent uncovered phrase

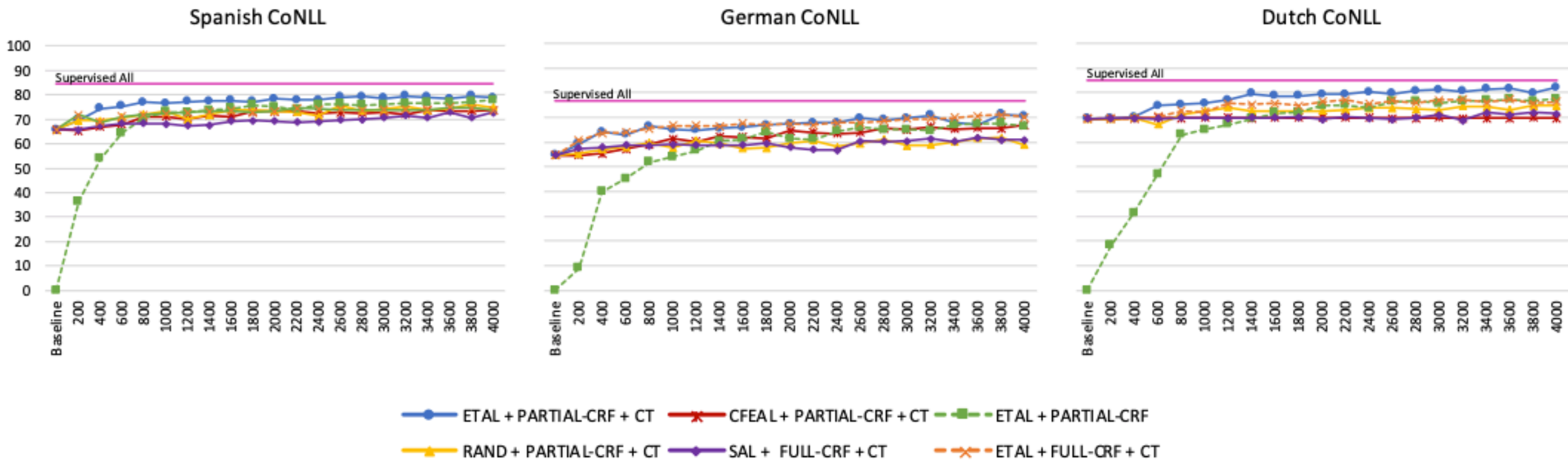


Zeng, Xiangkai, et al. "Empirical Evaluation of Active Learning Techniques for Neural MT." *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. 2019.

Bloodgood, Michael, and Chris Callison-Burch. "Bucking the trend: Large-scale cost-focused active learning for statistical machine translation." *arXiv preprint arXiv:1410.5877* (2014).

Cross-lingual Learning + Active Learning

- Both perform better than either in isolation



Chaudhary, Aditi, et al. "A little annotation does a lot of good: A study in bootstrapping low-resource named entity recognizers." *EMNLP 2019*.

Active Learning and Human Work

Human Effort and Active Learning

- In simulation, it's common to assess active learning based on words/sentences annotated

- However, in reality: active learning may select harder examples

  - Takes more time

  - Higher chance of human error

- Often simulations **over-estimate gain** from active learning

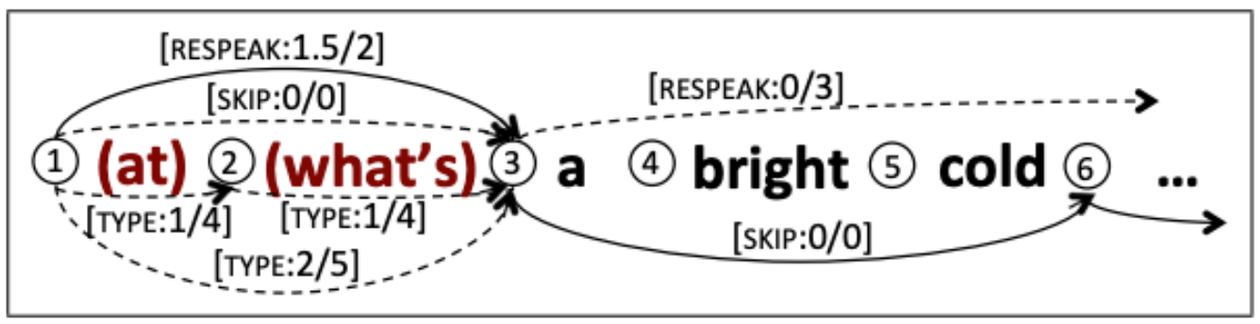
Considering Cost in Active Learning

- **Proactive learning:** considers different oracles that cost different amounts for each

- **Cost-sensitive Annotation:**

    - Create a model of annotation cost and accuracy gain for each span (in different modes)

    - Choose best spans/modes based on this

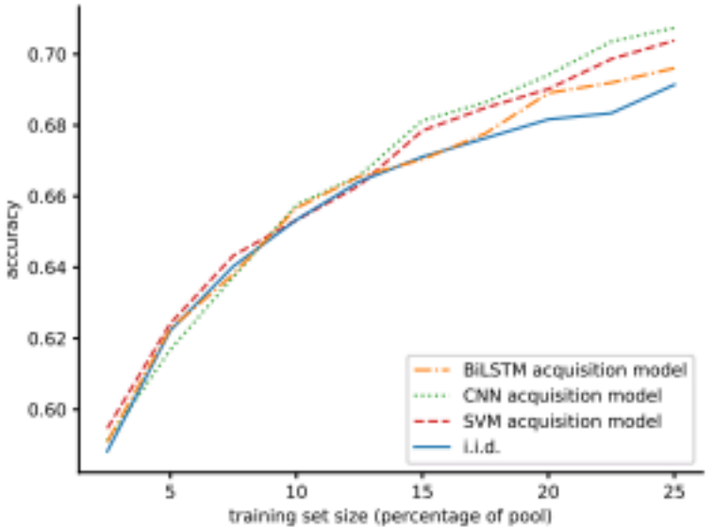


Donmez, Pinar, and Jaime G. Carbonell. "Proactive learning: cost-sensitive active learning with multiple imperfect oracles." *Proceedings of the 17th ACM conference on Information and knowledge management*. 2008.

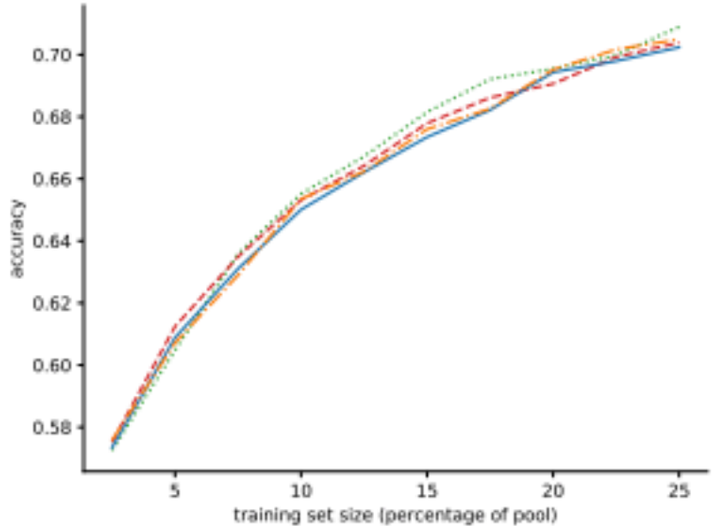
Sperber, Matthias, et al. "Segmentation for efficient supervised language annotation with an explicit cost-utility tradeoff." *Transactions of the Association for Computational Linguistics* 2 (2014): 169-180.

Reusability of Active Learning Annotations

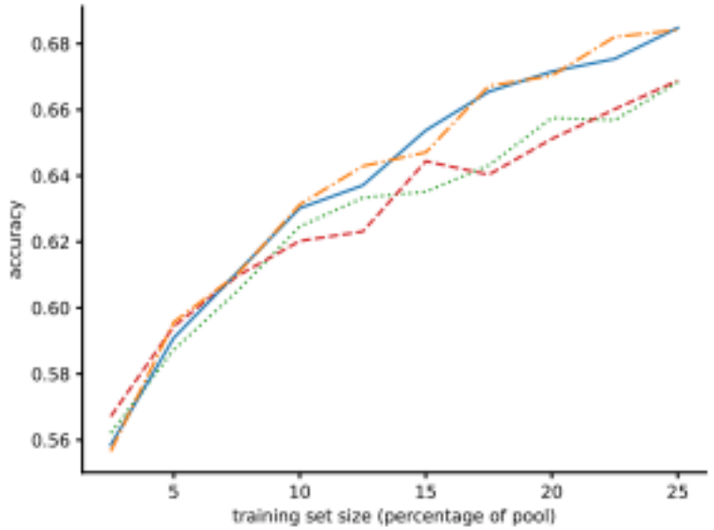
- If active learning annotations are obtained with one model, they may not transfer well to other models



(a) SVM on Movies dataset



(b) CNN on Movies dataset



(c) LSTM on Movies dataset

Lowell, David, Zachary C. Lipton, and Byron C. Wallace. "Practical obstacles to deploying active learning." *arXiv preprint arXiv:1807.04801* (2018).

Discussion Question

Discussion Question

- Given the task and language(s) you are tackling in your project, how could you use active learning improve?
    - How would you calculate uncertainty?
    - How would you calculate representativeness?
    - How would you ensure annotator productivity?