EMNLP | 2020

Georgia Institute of Technology

SeqMix: Augmenting Active Sequence Labeling via Sequence Mixup

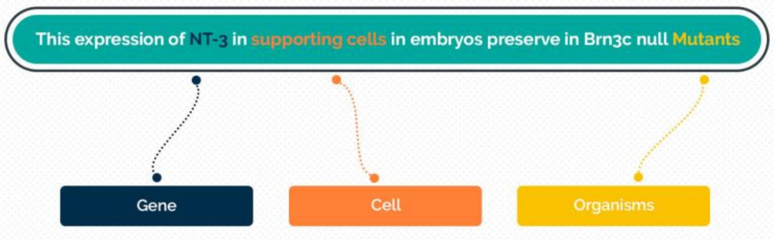
Rongzhi Zhang, Yue Yu, Chao Zhang

Georgia Institute of Technology

CS 7641/4641 Seminar | Sep 23, 2020

Introduction

- Sequence labeling is core to many NLP tasks
  - Part-of-speech (POS) tagging
  - Event extraction
  - Named entity recognition (NER)



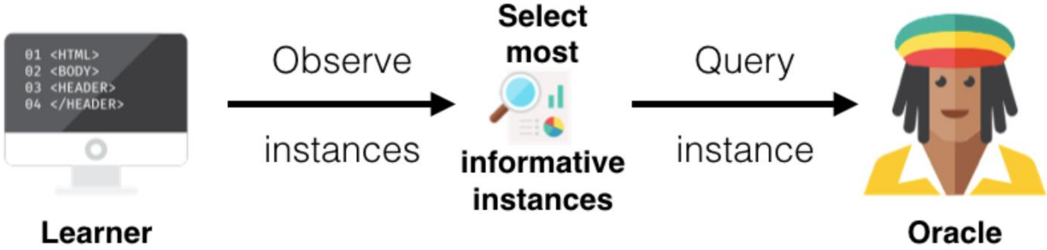
This expression of NT-3 in supporting cells in embryos preserve in Brn3c null Mutants

- Deep neural architectures have demonstrated superior performance for this task but they are label hungry.

Active Learning

Traditional methods: randomly sample a large dataset to train a model

Active learning: choose the data of interest in low-resource scenarios



Learner

Observe

instances

Select most informative instances

Query

instance

Oracle

However, existing methods on active sequence labeling use queried data samples alone in each iteration.

- The queried samples provide limited data diversity
- Using them alone is an inefficient way of leveraging annotation

We thus want to enhance active sequence labeling via data augmentation.

Challenges

We need to jointly generate sentences and token-level labels.

- Prevailing generative models are inapplicable

  -- They can only generate word sequences without labels.

- Heuristic data augmentation methods such as context-based words substitution, synonym replacement are also infeasible.

  -- Label composition is complex for sequence labeling. Directly manipulating tokens as above may inject incorrectly labeled sequences into training data.

Our Solution

- SeqMix searches for pairs of eligible sequences and mixes them both in the feature space and the label space
  - Implement linear interpolation in the embedding space.
  - Generate the sequences along with the labels.

- Deploy a discriminator to judge if the generated sequence is plausible or not
  - Compute the perplexity scores for all the generated sequences
  - Select the low-perplexity sequences as plausible ones

Problem Definition

- Traditional active learning starts from a small labeled seed set L, and update it with the newly labeled query samples $\langle X, Y \rangle$ in each learning round as $L = L \cup \langle X, Y \rangle$

- Formally, we define our task as:
  1) Construct a generator $\phi(\cdot)$ to implement sequence and label generation based on the actively sampled data X and its label Y
  2) Set a discriminator $d(\cdot)$ to yield the filtered generation
  3) Augment the labeled set as $L = L \cup \langle X, Y \rangle \cup d(\phi(X, Y))$

Active Learning for Sequence Labeling

- Active sequence labeling selects K most informative instances in each learning round. The representative query polices to measure the informativeness are as below.
- Least Confidence (LC)

$$\gamma^{\mathrm{LC}}(\mathbf{x}) = 1 - \max_{y^*}(P(\mathbf{y}^*|\mathbf{x};\theta))$$

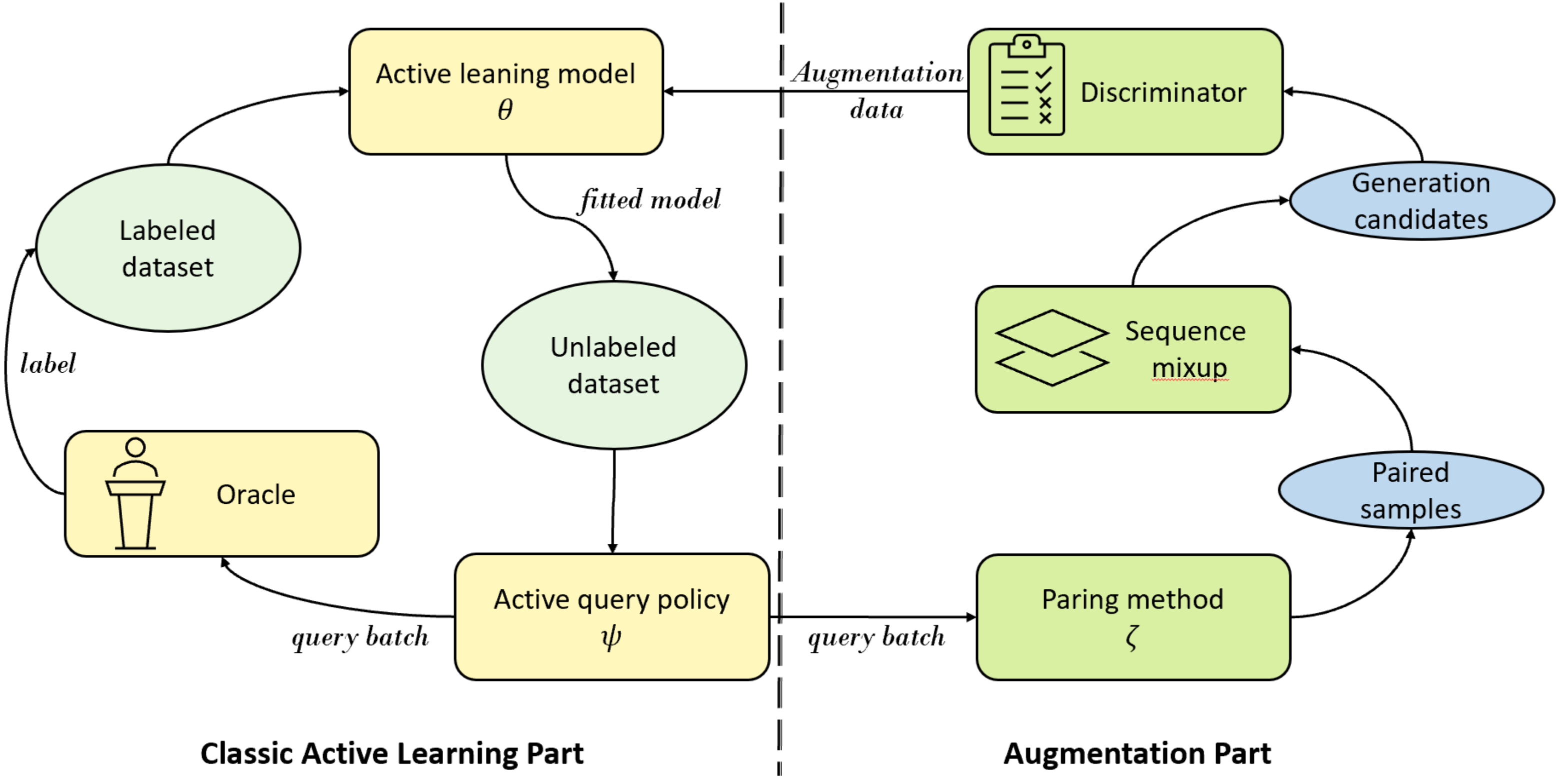
- Normalized Token Entropy (NTE)

$$\gamma^{\mathrm{TE}}(\mathbf{x}) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M} P_m(\mathbf{y}_t|\mathbf{x},\theta)\log P_m(\mathbf{y}_t|\mathbf{x},\theta)$$

- Disagreement Sampling

$$\gamma^{\mathrm{VE}}(\mathbf{x}) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M} \frac{V_m(\mathbf{y}_t)}{C}\log\frac{V_m(\mathbf{y}_t)}{C}$$

Method Overview



Sequence Mixup in the Embedding Space

- Given two input samples x_i, x_j along with their labels y_i, y_j,

  The mixing process is:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j,$$
$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j,$$

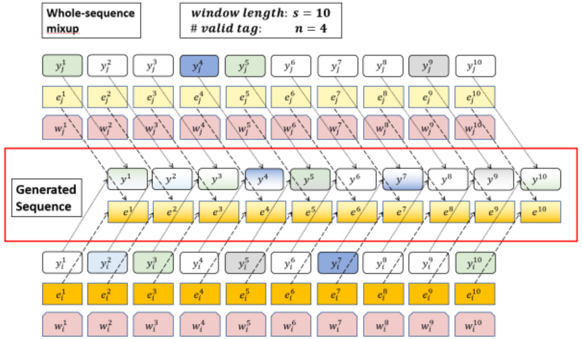
  where $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$.

- The input space is discrete for text, so we make linear interpolation in the embedding space.

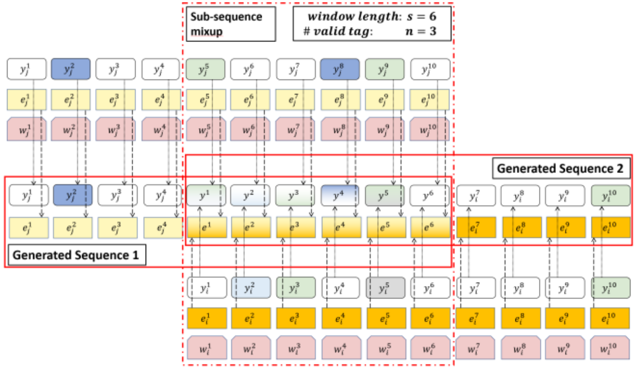
  For the embedding and label of the t-th token in a sequence:

$$\mathbf{e}^t = \underset{\mathbf{e} \in \mathcal{E}}{\arg\min} \left\| \mathbf{e} - (\lambda \mathbf{e}_\mathrm{i}^\mathrm{t} + (1 - \lambda) \mathbf{e}_\mathrm{j}^\mathrm{t}) \right\|_2$$
$$\mathbf{y}^t = \lambda \mathbf{y}_i^t + (1 - \lambda) \mathbf{y}_j^t$$

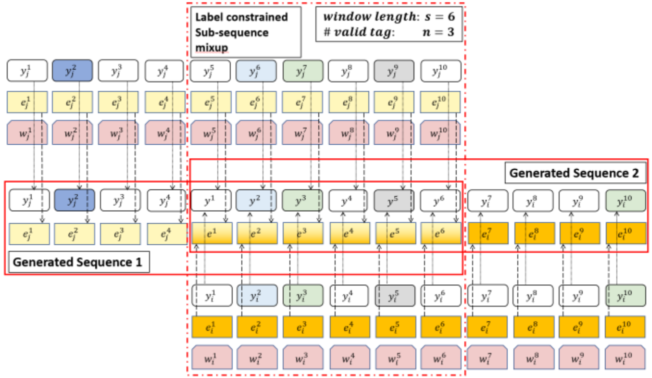
Multi-granularity Sequence Mixup



1. Whole-sequence mixup



2. Sub-sequence mixup



3. Label-constrained sub-sequence mixup

Scoring Mechanism

- The generation quality fluctuates due to the following reasons:
  - The mixing coefficient λ sampled from $Beta(\alpha, \alpha)$ determines the interpolation strength.
  - The discrete embedding space can hardly match a mixed embedding exactly.

- To maintain the quality of mixed sequences , we set a discriminator to score the perplexity of the sequences.
  - Utilize a language model to score the sequence X by computing its perplexity

$$\mathrm{Perplexity}(\mathbf{x}) = 2^{-\frac{1}{\mathrm{T}}\sum_{\mathrm{i}=1}^{\mathrm{T}} \log \mathrm{p}(\mathrm{w}_\mathrm{i})}$$

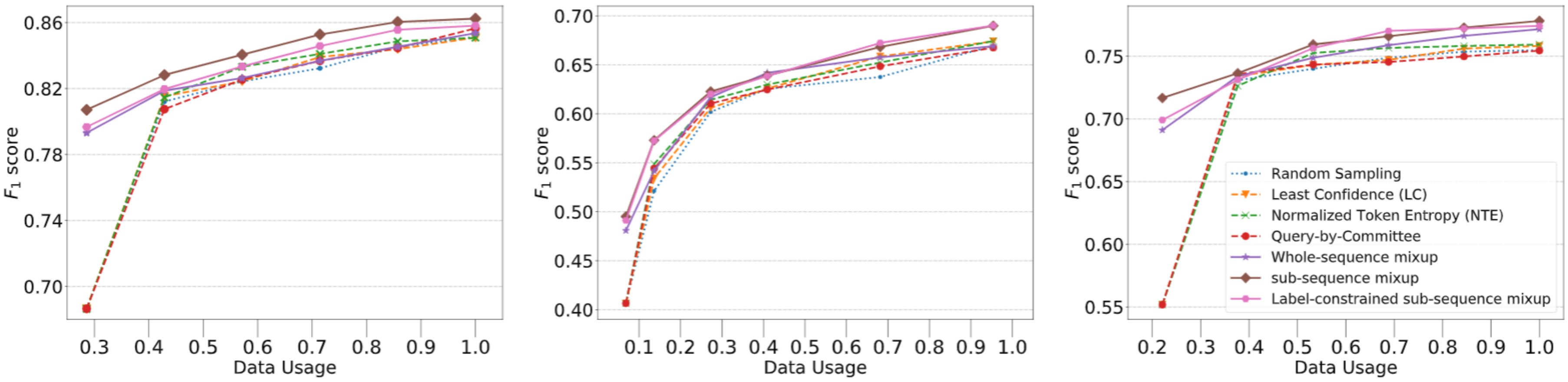
  - Based on the perplexity and a score range $[s_1, s_2]$, give judgement for the sequence X

$$d(\mathbf{x}) = \mathbb{1}\left\{s_1 \le \mathrm{Perplexity}\left(\mathbf{x}\right) \le \mathrm{s}_2\right\}$$

Experiments

- Datasets
  - CoNLL-03 -- a well studied dataset for NER task
  - ACE-05 -- a well-known corpus for automatic content extraction
  - WebPage – a tiny NER corpus comprise of 20 webpages
- Baseline
  - Random sampling
  - Least confidence sampling
  - Normalized Token Entropy sampling
  - Query-by-Committee sampling
- Evaluation
  - Set 6 data usage percentiles for the training set, evaluate F_1 score for each data usage percentile.

Main Results



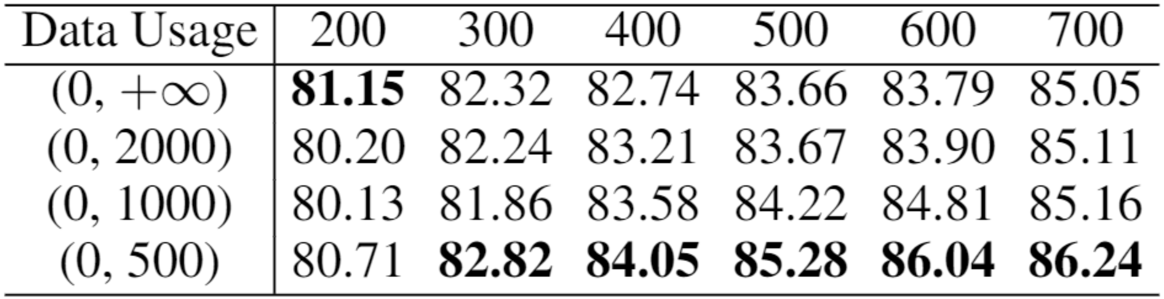
(a) CoNLL 2003 (700 labeled data)

(b) ACE05 (14k labeled data)

(c) WebPage (385 labeled data)

- SeqMix consistently outperforms the baselines at each data usage percentile.
- The augmentation advantage is especially prominent for the seed set initialization stage where the annotation is very limited.

Ablation Study: Effect of Discriminator



Data Usage	200	300	400	500	600	700
$(0, +\infty)$	**81.15**	82.32	82.74	83.66	83.79	85.05
(0, 2000)	80.20	82.24	83.21	83.67	83.90	85.11
(0, 1000)	80.13	81.86	83.58	84.22	84.81	85.16
(0, 500)	80.71	**82.82**	**84.05**	**85.28**	**86.04**	**86.24**

The performance of SeqMix with variant discriminator score range

- The score range $(0, +\infty)$ indicates no discriminator participated.
- The comparison demonstrates the lower the perplexity, the better the generation quality.
- The score range can further narrow down, but we made a trade-off between the generation quality and the generation quantity.

Summary

- We propose SeqMix to augment active sequence labeling
  - Introduce data diversity through the sequence Mixup in latent space
  - Alleviate the dependency to the annotation capacity for active learning.

- Future Work
  - We plan to explore implementing SeqMix via the combination of a multi-layer representation.
  - We are also interested in combining SeqMix with other active learning method and extending SeqMix to other NLP tasks.

Thank you!

Questions?