

EMNLP | 2020



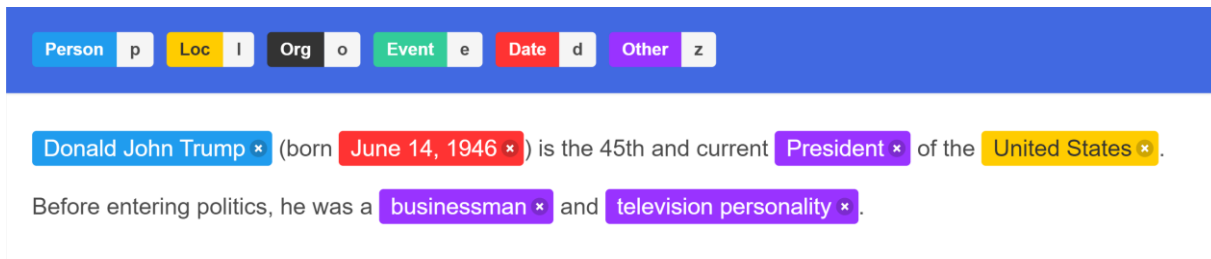
SeqMix: Augmenting Active Sequence Labeling via Sequence Mixup

Rongzhi Zhang, Yue Yu, Chao Zhang

Georgia Institute of Technology

Introduction

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

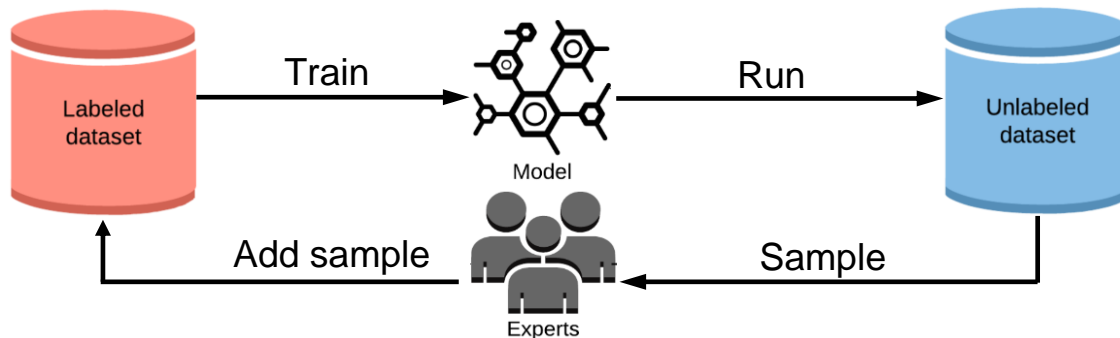


The image shows a screenshot of a Named Entity Recognition (NER) interface. At the top, there is a legend bar with six categories: Person (p), Loc (l), Org (o), Event (e), Date (d), and Other (z). Below the legend, a text snippet is displayed with several entities highlighted in colored boxes: "Donald John Trump" (Person, blue), "June 14, 1946" (Date, red), "President" (Other, purple), "United States" (Loc, yellow), "businessman" (Other, purple), and "television personality" (Other, purple). Each highlighted entity has a small 'x' icon to its right.

- Neural sequential models have shown strong performance for sequence labeling but they are **label hungry**.

Active Sequence labeling

- Active learning is suitable for sequence labeling in **low-resource scenarios**.



- 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 study the problem of **enhancing active sequence labeling via data augmentation**.

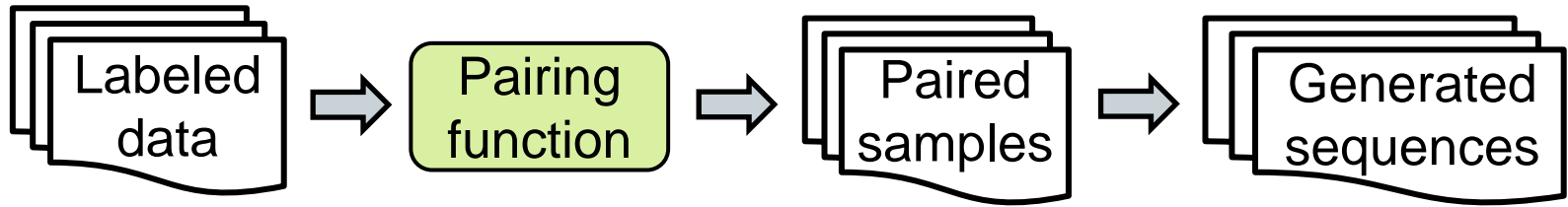
Challenges

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

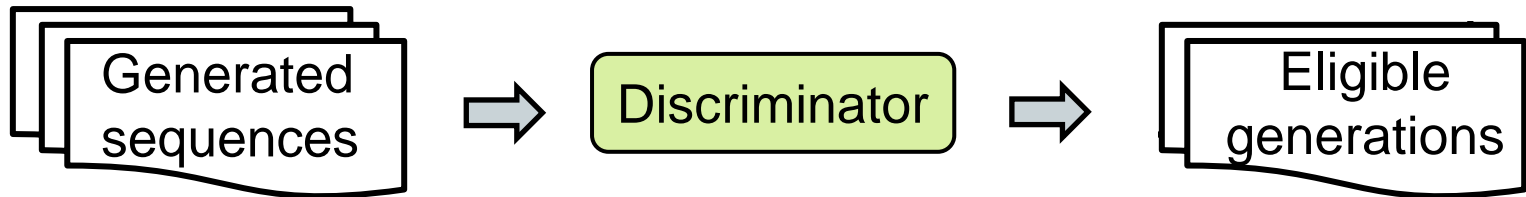
- Prevailing generative models are inapplicable.
 - They can only generate word sequences **without labels**.
- Heuristic data augmentation methods are infeasible.
 - Directly manipulating tokens such as *context-based words substitution*, *synonym replacement* may **inject incorrectly labeled sequences** into training data.

Our Solution

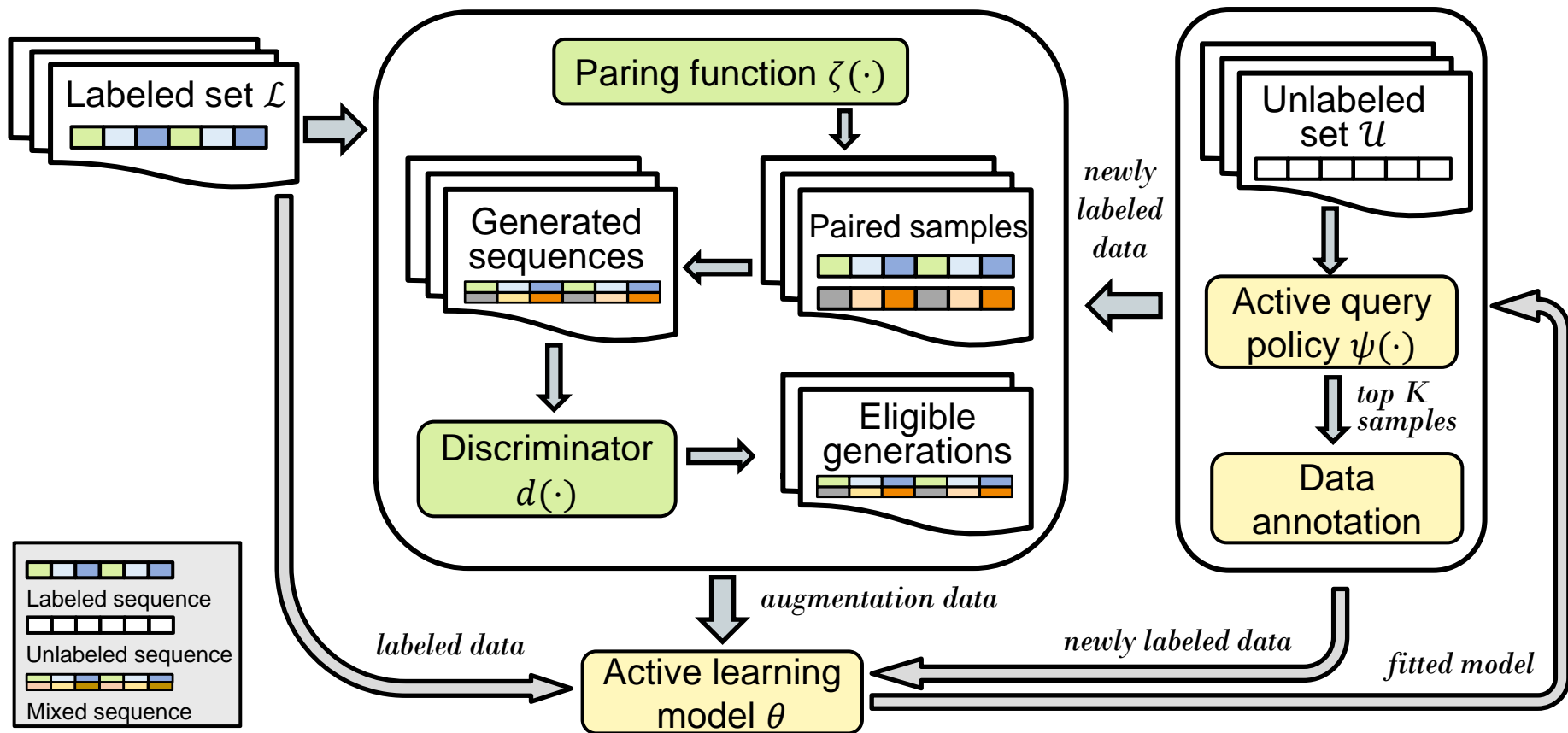
- SeqMix searches for **pairs of eligible sequences** and **mixes** them both in the feature space and the label space.



- Deploy a **discriminator** to judge if the generated sequence is plausible or not.

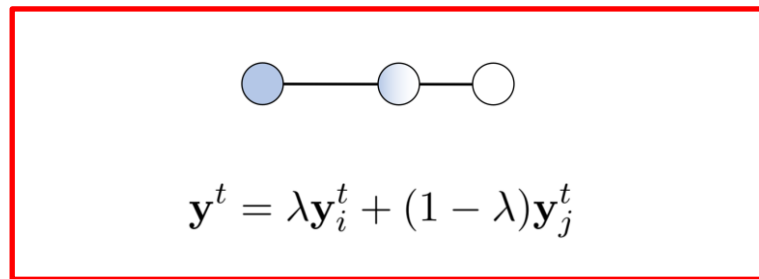
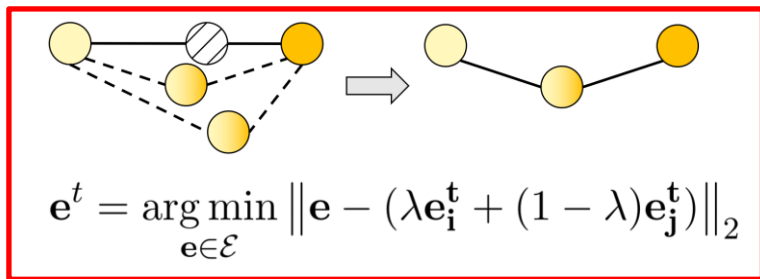


Method Overview



Sequence Mixup in the Embedding Space

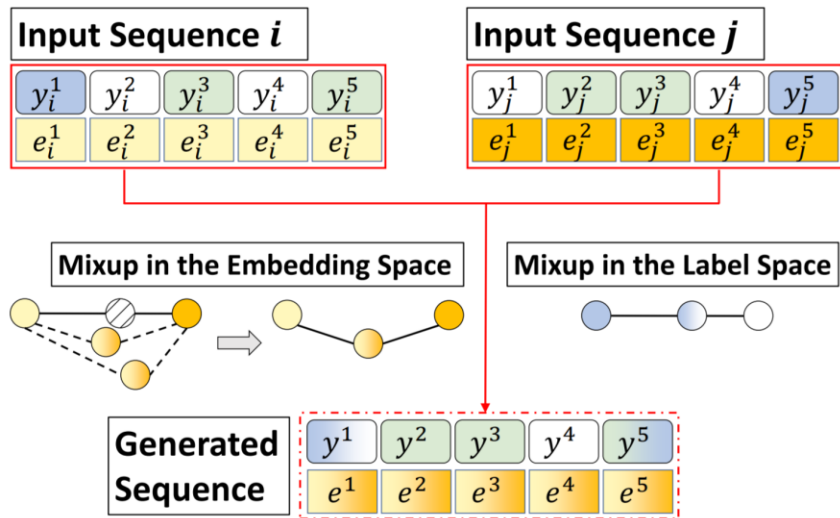
- The **input space is discrete** for text, so we make linear interpolation in the **embedding space**.
- Given two sequences x_i and x_j , the mixing process at the t-th position:



where e^t is the mixed embedding, y^t is the mixed label, \mathcal{E} is the pre-defined embedding list and the mixing coefficient $\lambda \sim \text{Beta}(\alpha, \alpha)$.

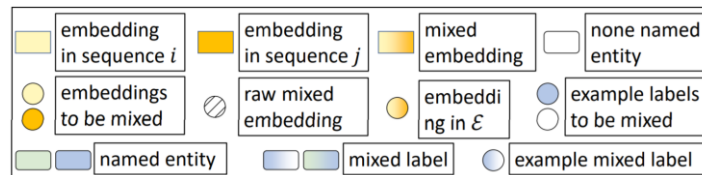
Whole-sequence Mixup

- Perform sequence mixing at the **whole-sequence level**.
- May include incompatible sub-sequences and generate implausible sequences.



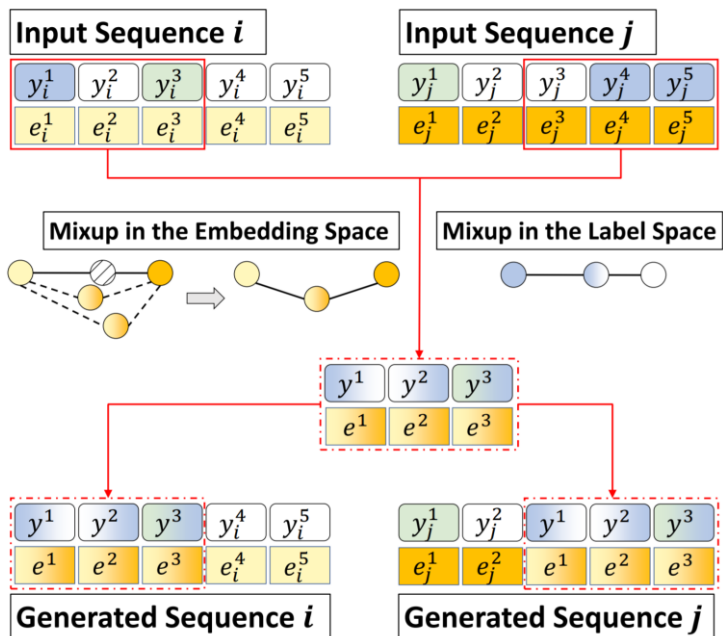
1. Sequence length $s = 5$, valid label density threshold $\eta_0 = \frac{3}{5}$.

2. Red solid frames indicates the whole sequences with **same length** and valid label density $\eta \geq \eta_0$ get paired.

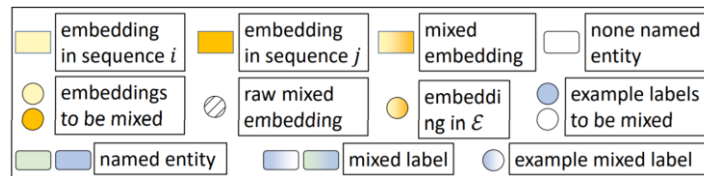


Sub-sequence Mixup

- Require the **sub-sequences** of two input sequence are paired.
- Keep the syntax structure** of the original sequence, while **providing data diversity**.

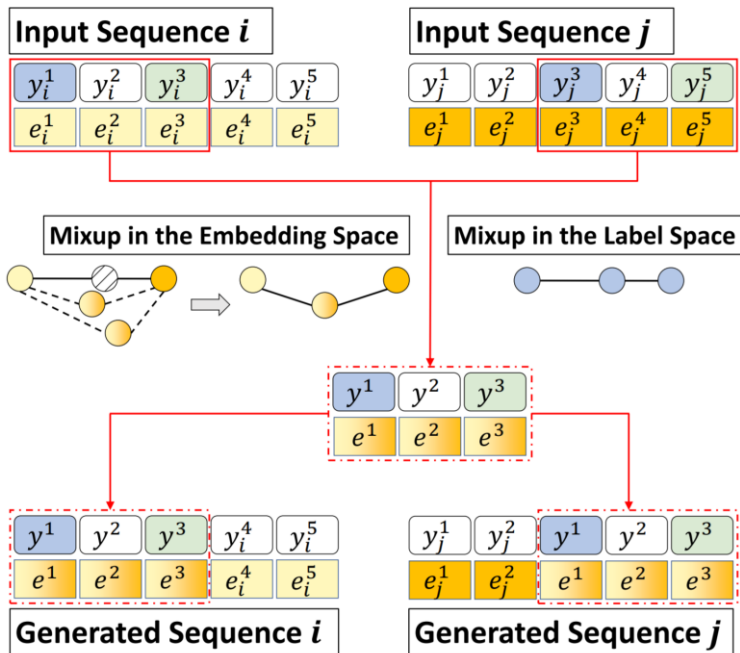


- Sub-sequence length $s = 3$, valid label density threshold $\eta_0 = \frac{2}{3}$.
- Red solid frames indicates the sub-sequences with **same length** and valid label density $\eta \geq \eta_0$ get paired.

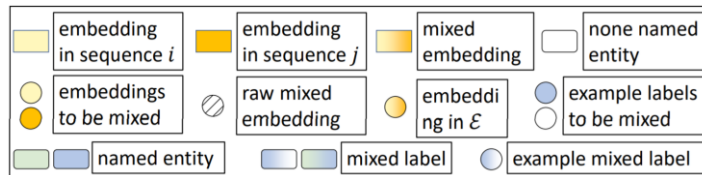


Label-constrained sub-sequence Mixup

- A **special case** of the sub-sequence mixup.
- Further require the **labels** of sub-sequences are **consistent**.



1. Sub-sequence length $s = 3$, valid label density threshold $\eta_0 = \frac{2}{3}$.
2. Red solid frames indicates the sub-sequences with **same length**, **consistent labels**, and valid label density $\eta \geq \eta_0$ get paired.



Scoring and Selecting Plausible Sequences

- 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.

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

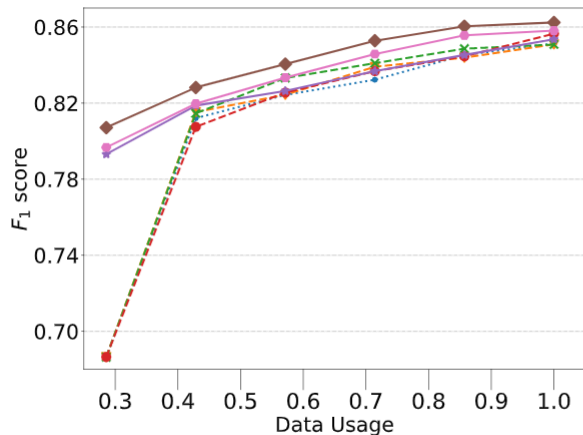
- Based on the perplexity and a score range $[s_1, s_2]$, give judgement for the sequence X .

$$d(\mathbf{x}) = \mathbb{1} \{s_1 \leq \text{Perplexity}(\mathbf{x}) \leq s_2\}$$

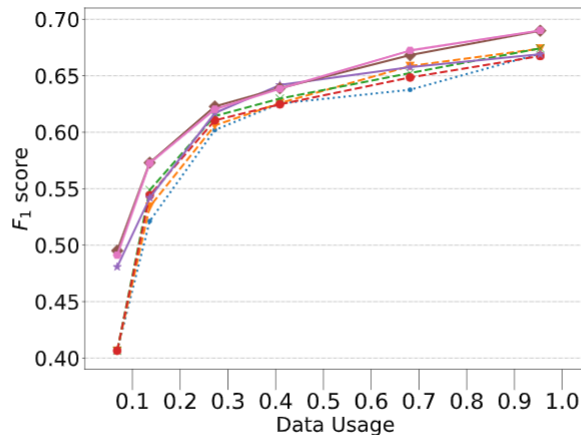
Experiments

- 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
 - 4 active learning methods
- Evaluation
 - Set 6 data usage percentiles for the training set, calculate F_1 score for each data usage percentile.

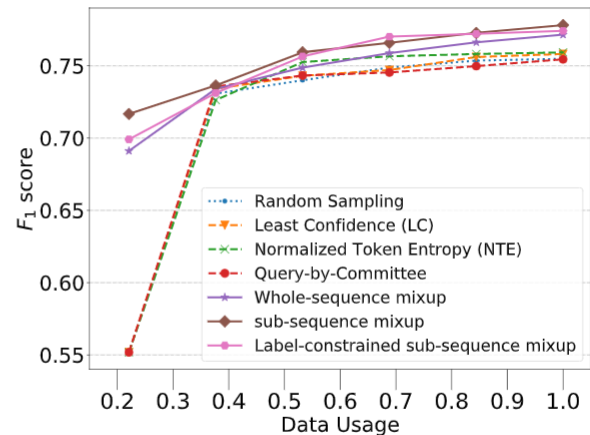
Main Results



(a) CoNLL 2003 (700 labeled data)



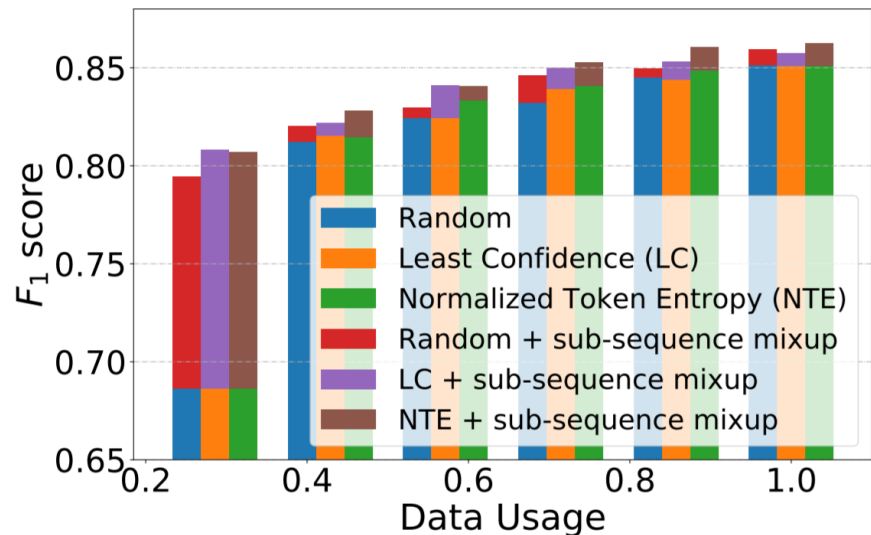
(b) ACE05 (14k labeled data)



(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.

Enhance different active learning policies



The improvements to different active learning approaches provided by SeqMix.

- SeqMix is **generic** to various active learning policies.
- For random sampling, LC sampling and NTE sampling, the averaged performance gain is {2.46%, 2.85%, 2.94%}.

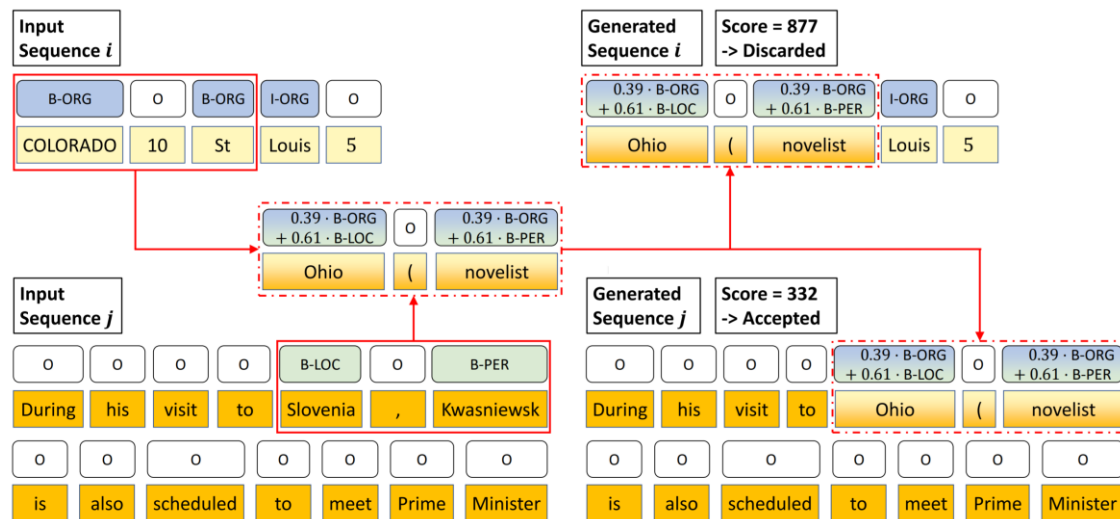
Ablation Study: Effect of Discriminator

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**.

Case Study: The Generation Process



- Sub-sequence length $s = 3$, valid label density threshold $\eta_0 = \frac{2}{3}$, the **perplexity score threshold is 500**.
- Generated sequence i with perplexity score **877** is **discarded**.
- Generated sequence j with perplexity score **332** is **accepted**.

Summary

- We propose a data augmentation method SeqMix to enhance active sequence labeling
 - **Data diversity** introduced via the sequence Mixup in latent space.
 - **Plausible** augmented sequences generated.
 - **Generic** to various active learning policies.
- Future Work
 - Implement SeqMix by using the combination of a multi-layer representation of language models.
 - Harness external knowledge for further improving the diversity and plausibility of the generated data.
- Code
 - <https://github.com/rz-zhang/SeqMix>