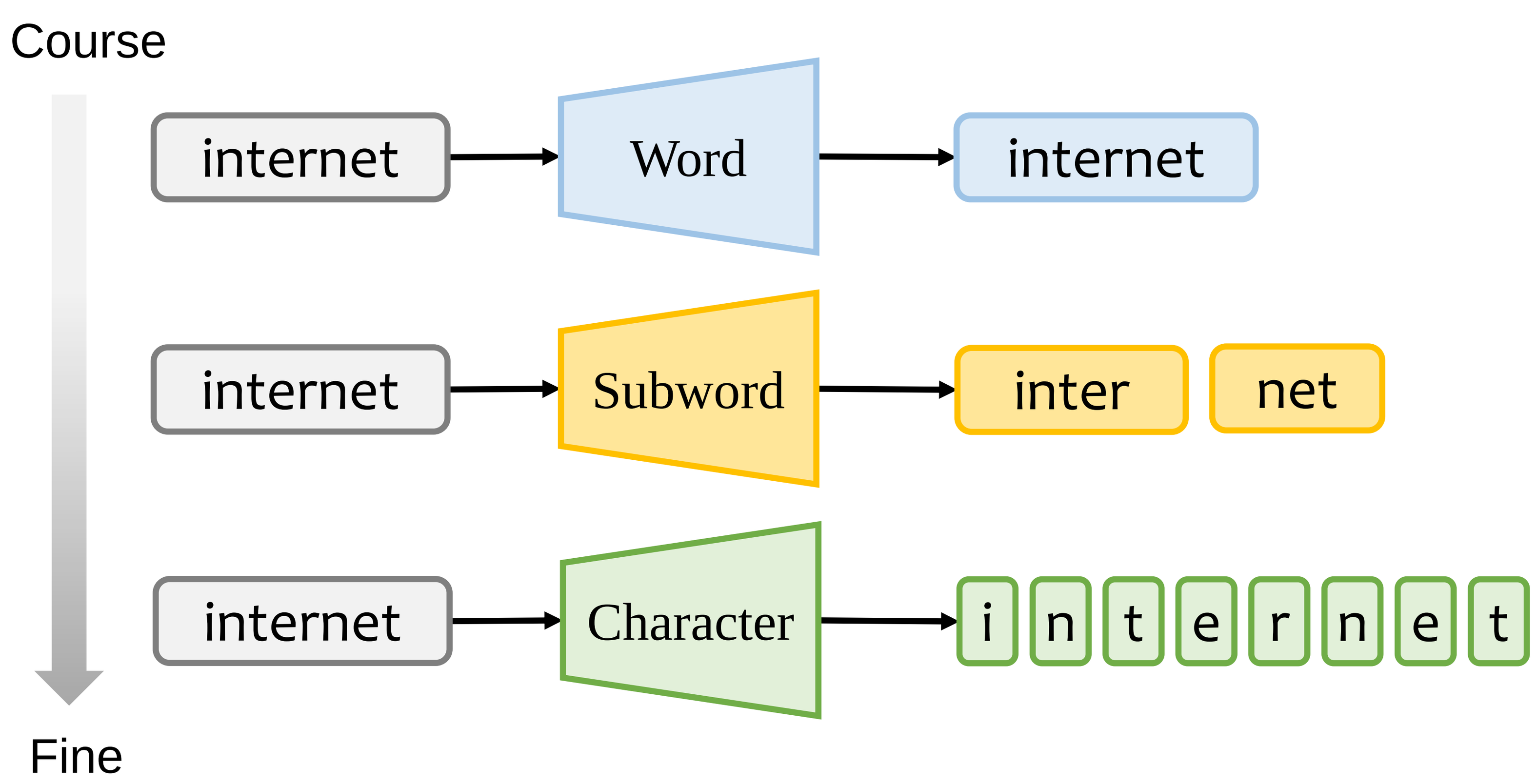
Progressive and Consistent Subword Regularization for Neural Machine Translation

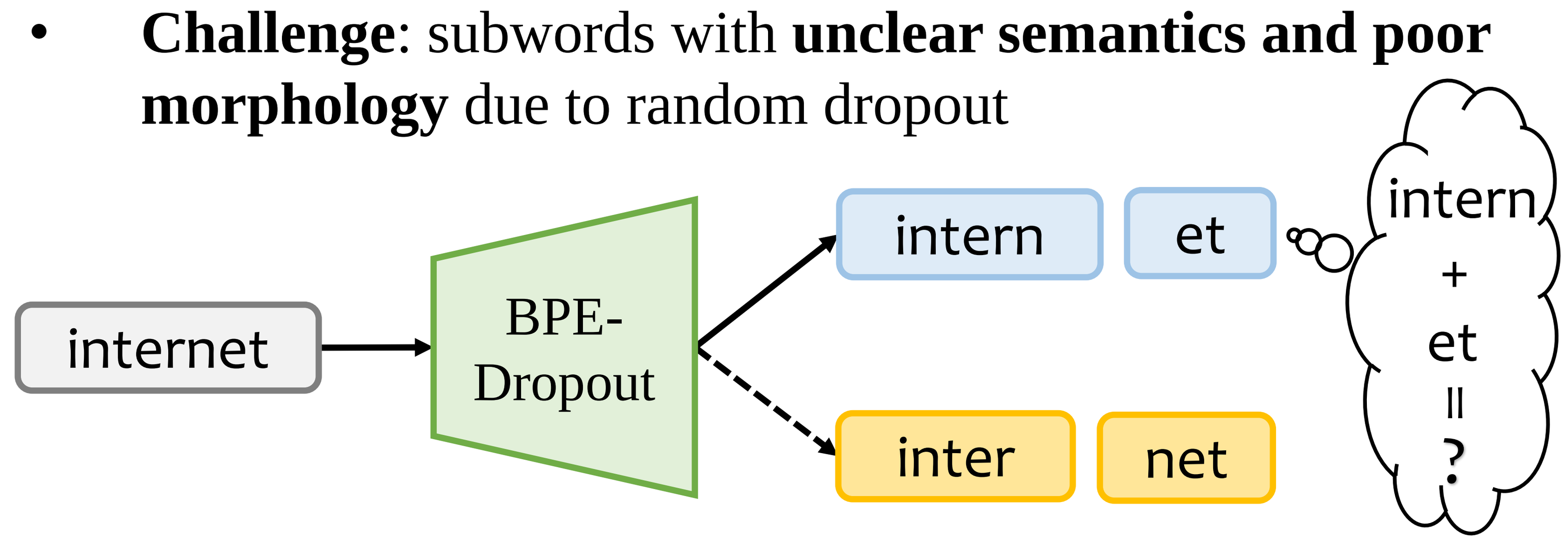
Yongqi Gao[1], Yingfeng Luo[1], Qinghong Zhang[1], Huibo Shao[1], Tong Xiao[1,2] * and Jingbo Zhu[1,2]
NLP Lab, School of Computer Science and Engineering, Northeastern University[1]
NiuTrans Research[2]

Motivation

- Tokenization
  - Tokenizers with different granularity
  - **Challenge: unique segmentation**

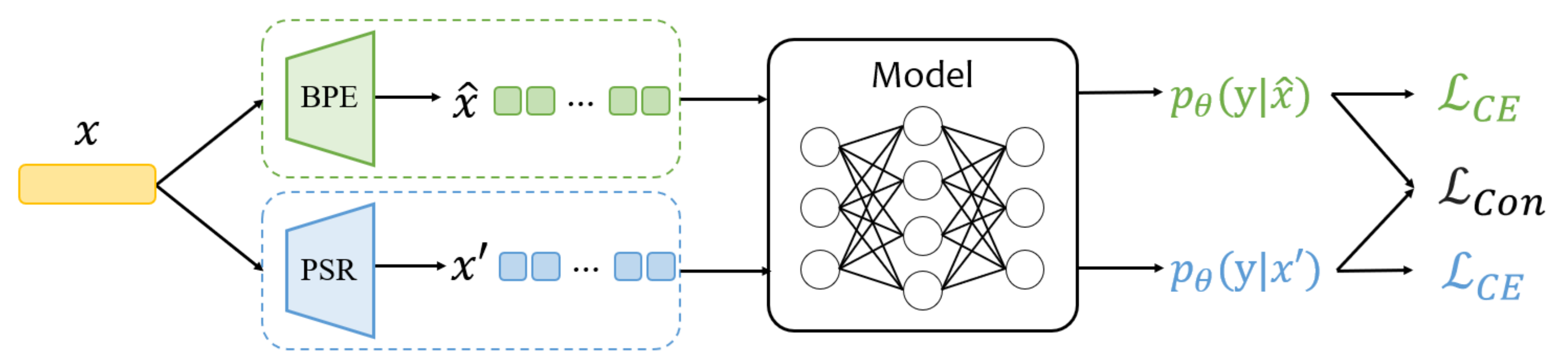


- Subword Regularization
  - BPE-Dropout: multiple subwords from random dropout
  - **Challenge**: subwords with **unclear semantics and poor morphology** due to random dropout



Progressive and Consistent Subword Regularization

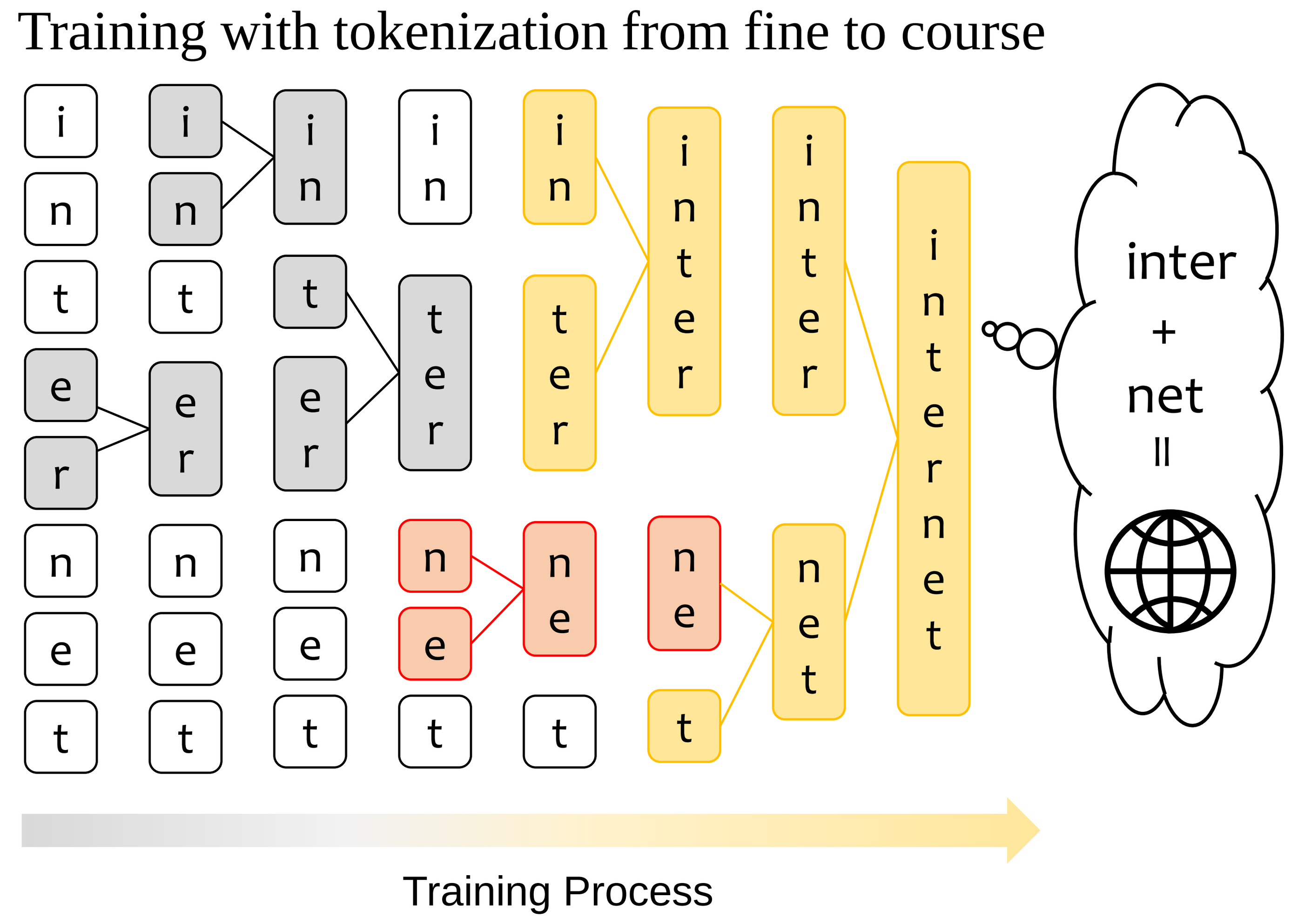
- Consistent Subword Regularization
  - $\mathcal{L}_{CE}$: cross-entropy loss for each segmentation
  - $\mathcal{L}_{CON}$: distance between each prediction distribution



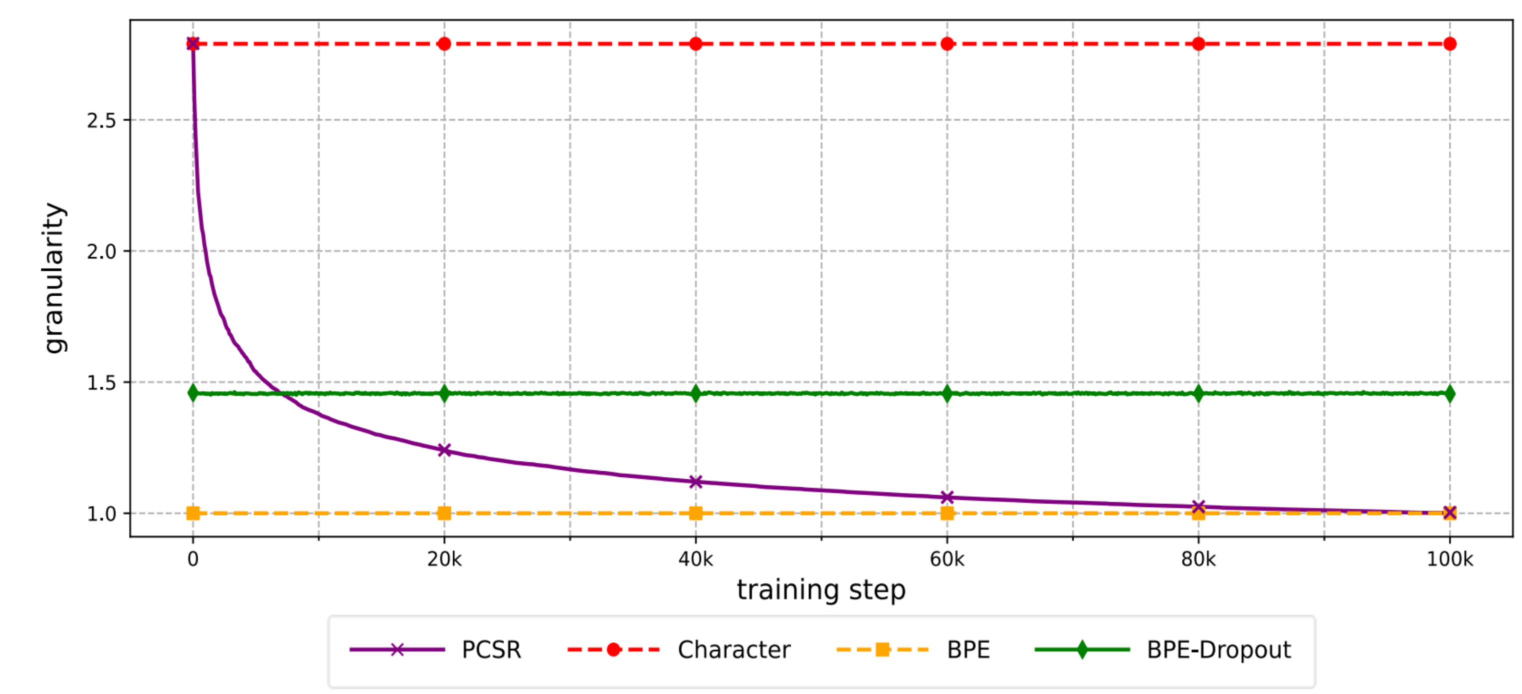
- $\mathcal{L}$: total loss balancing $\mathcal{L}_{CE}$ and $\mathcal{L}_{CON}$ by λ, forcing the outputs to be **accurate and consistent**

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{CON}$$

- Progressive Subword Regularization
  - Training with tokenization from fine to course



- Granularity variation during training



Main Results

- Naive Subword Regularization **vs** Baseline
  - BPEDrop and PSR > BPE and Character
- Consistent Subword Regularization **vs** Naive Regularization
  - Tokenization+ $\mathcal{L}_{CON}$ > Tokenization itself
- PCSR (Our Method) **vs** Consistent BPE-Dropout
  - PCSR > Consistent BPE-Dropout

Methods	IWSLT14 DE→EN	IWSLT14 EN→DE	IWSLT17 FR→EN	IWSLT17 EN→FR	WMT14 EN→DE	WMT16 EN→RO
BPE	34.3	29.1	36.6	36.3	28.0	33.6
Character	32.7	27.6	36.0	36.3	27.1	31.9
BPEDrop	34.8	29.0	37.4	37.6	28.1	34.2
PSR	34.5	29.3	37.1	37.2	27.9	33.9
R-Drop	36.2	30.7	**38.3**	38.1	28.7	35.3
Character+$\mathcal{L}_{CON}$	35.6	29.8	37.7	37.7	28.1	34.4
BPEDrop+$\mathcal{L}_{CON}$	**36.7**	30.7	38.1	37.9	28.3	35.2
PCSR	**36.7**	**30.9**	**38.3**	**38.6**	**28.9**	**35.5**

Analysis

- Properties of Learned Embeddings
  - Nearest embedding neighbors

appoint PCSR	appoint BPE	similar PCSR	similar BPE	withdraw PCSR	withdraw BPE	invite PCSR	invite BPE
appointing	wledge	Similarly	simil@@	withdrawn	wledge	invites	invites
appoin@@	社	comparable	ก	withdrawal	ı	invitation	pite
adjust@@	》	simil@@	y-to-day	withdrawal@@	ˆ	inviting	ก

  - Short embeddings distance between the rare and the common
- Training on Agglutinative Languages
- Robustness to Out-of-domain Input

Methods	FLORES SI→EN	WMT18 TR→EN	Multi-domain IT	Koran	Law	Medical	Subtitles
BPE	6.5	19.1	14.1	8.9	27.4	24.6	14.9
BPEDrop	6.9	18.8	15.2	10.2	30.1	25.6	16.4
PSR	6.6	18.9	14.4	9.8	29.7	25.0	16.3
R-Drop	8.5	20.7	15.3	9.5	29.1	27.1	16.1
BPEDrop+$\mathcal{L}_{CON}$	**8.6**	20.7	15.3	**10.8**	30.5	27.0	18.1
PCSR	**8.6**	**21.0**	**15.5**	10.0	**30.9**	**27.3**	**18.2**

Conclusion

✓ We propose PCSR, a simple subword regularization method based on **progressive granularity**.
✓ We highlight the critical role of **consistency constraints** in subword regulation for NMT.
✓ Future research could apply PCSR to **other NLP tasks**.

2024 THE 13TH CCF INTERNATIONAL CONFERENCE ON NATURAL LANGUAGE PROCESSING AND CHINESE COMPUTING