Carnegie Mellon University
Language Technologies Institute

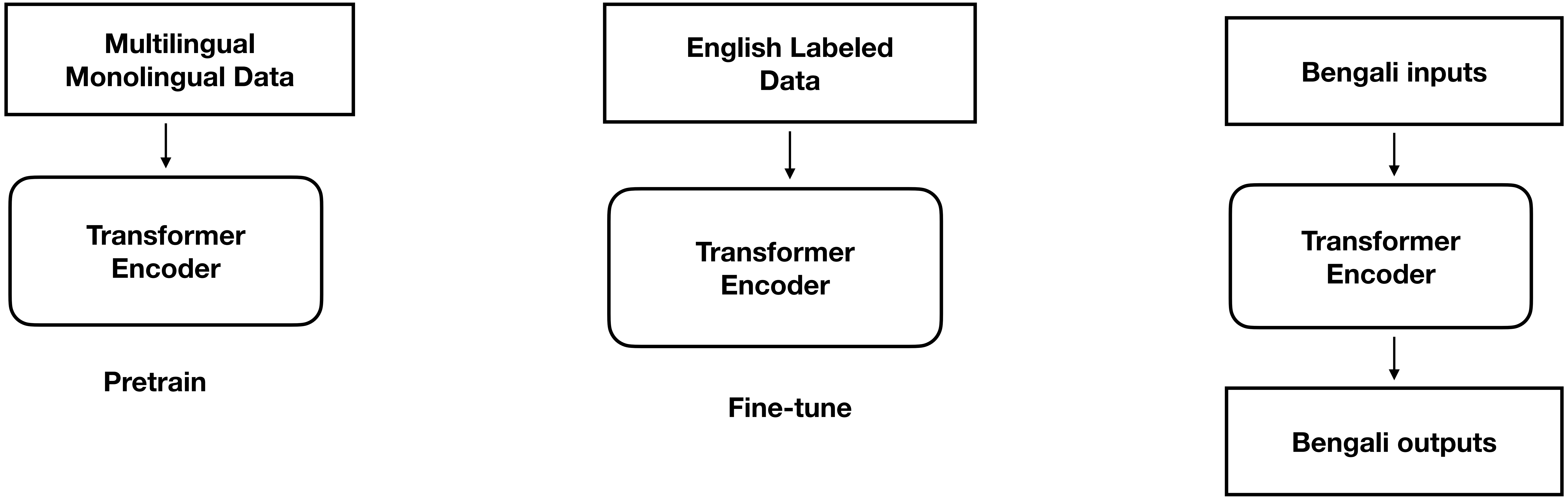
DeepMind

Multi-view Subword Regularization

Xinyi Wang[1]
Sebastian Ruder[2]
Graham Neubig[1]

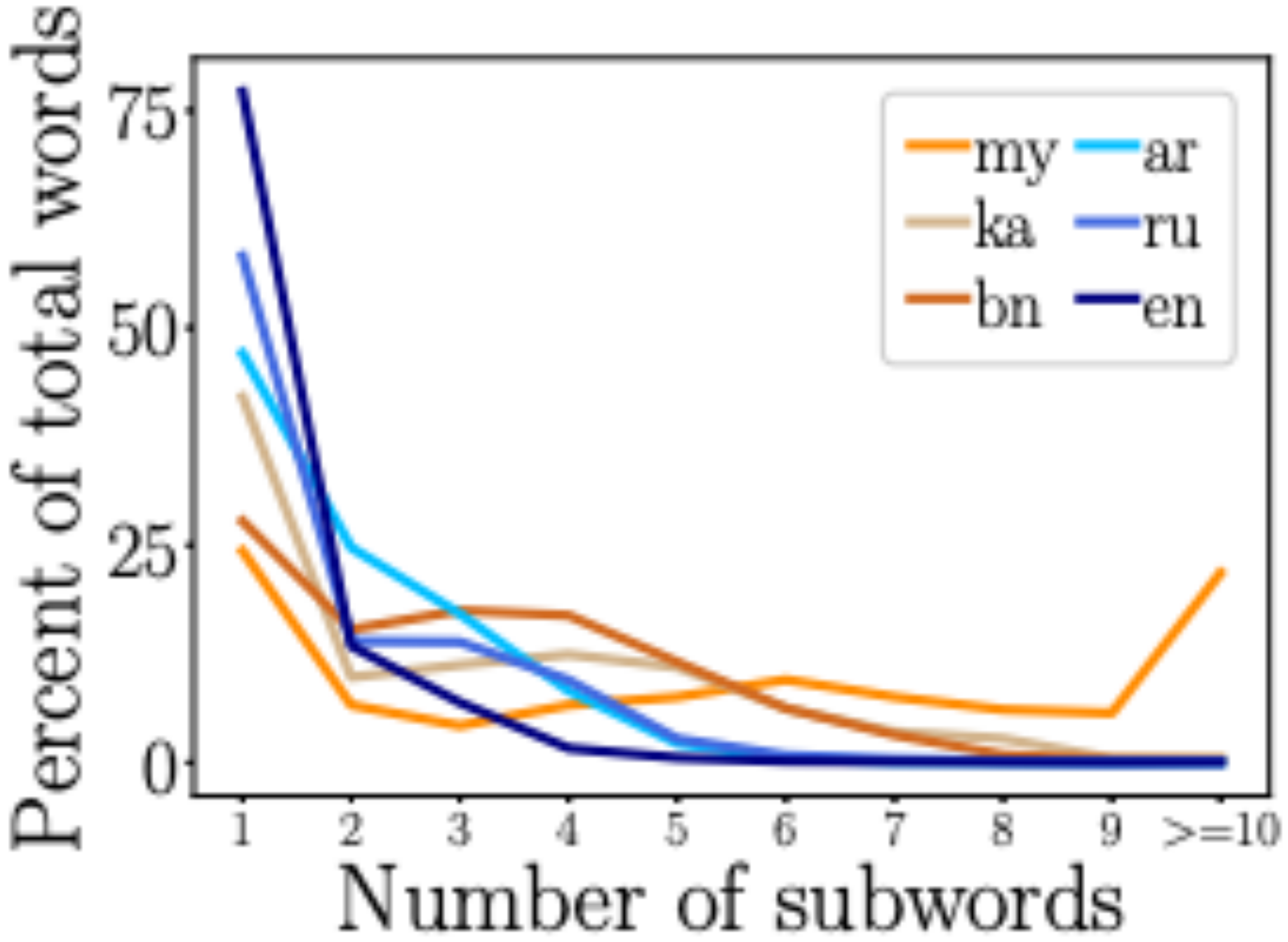
1. Language Technologies Institute, CMU
2. DeepMind

Multilingual Pretrained Models



- **Zero-shot cross-lingual transfer**: fine-tune model on English, generalize to other languages
- Utilize a single subword vocabulary constructed from monolingual data in hundreds of languages
- These models suffer from **suboptimal subword segmentation**

Subword Segmentation is Suboptimal



- Many low-resource languages tend to be over-segmented

Subword Segmentation is Suboptimal

en	excitement	**fr**	excita/tion
de	Auf/re/gung	**pt**	excita/ção
el	εν/ϑ/ουσι/ασμός	**ru**	волн/ение

Table. XLM-R segmentation of “excitement” in different languages

- Mismatch in segmentation could harm cross-lingual transfer

Subword Segmentation is Suboptimal

- Existing methods
  - Embed words using characters (Ma et. al. 2020)
  - Separately construct subword segmentation for each language cluster (Chung et. al. 2020)
  - Add a phrase-level segmentation (Zhang et. al. 2020)
- Modifying subword vocabulary requires retraining the large language model
- What is a **computationally efficient approach** for this problem at **fine-tuning** time?

Background: Subword Segmentation

Always segment **Excitement -> Excite/ment**

- Deterministic segmentation
  - Byte-pair encoding (BPE; Sennrich et. al. 2016)
  - Unigram language model (ULM; Kudo et. al. 2018)

Background: Subword Segmentation

Samples from segments **Excitement -> Excitement**
-> Excite/ment
-> Exc/ite/ment

- Probabilistic segmentation
    - BPE-dropout (Provikov et. al. 2020)
    - ULM-sample (Kudo et. al. 2018)

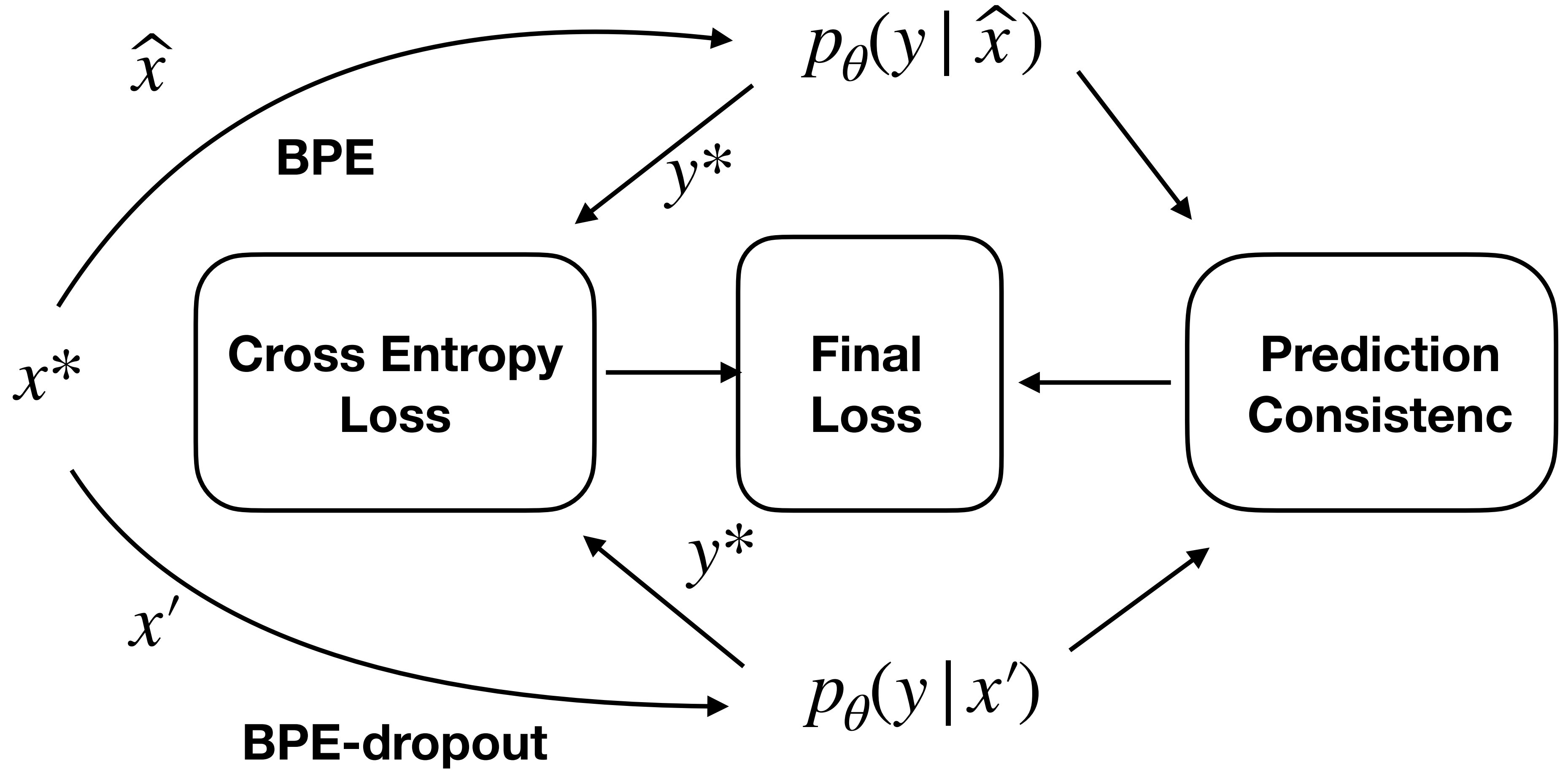
Background: Subword Regularization

- Simply use probabilistic segmentation during training time
- Has only been applied in NMT to improve model performance and robustness

Subword Regularization for Cross-lingual Transfer

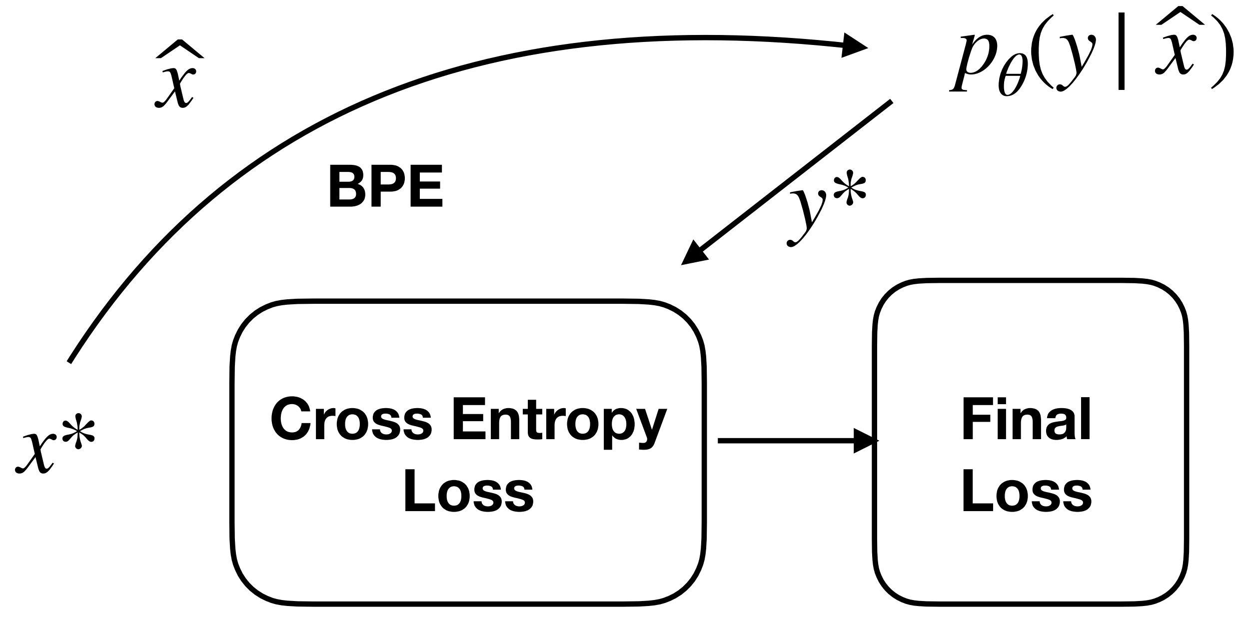
- We propose to use SR at **fine-tuning** time of multilingual pertained models
- It's a simple method but could make the model more accommodating to segmentation disparities in different languages
- However, might cause **segmentation discrepancy between pretraining and fine-tuning**

Multi-view Subword Regularization (MVR)



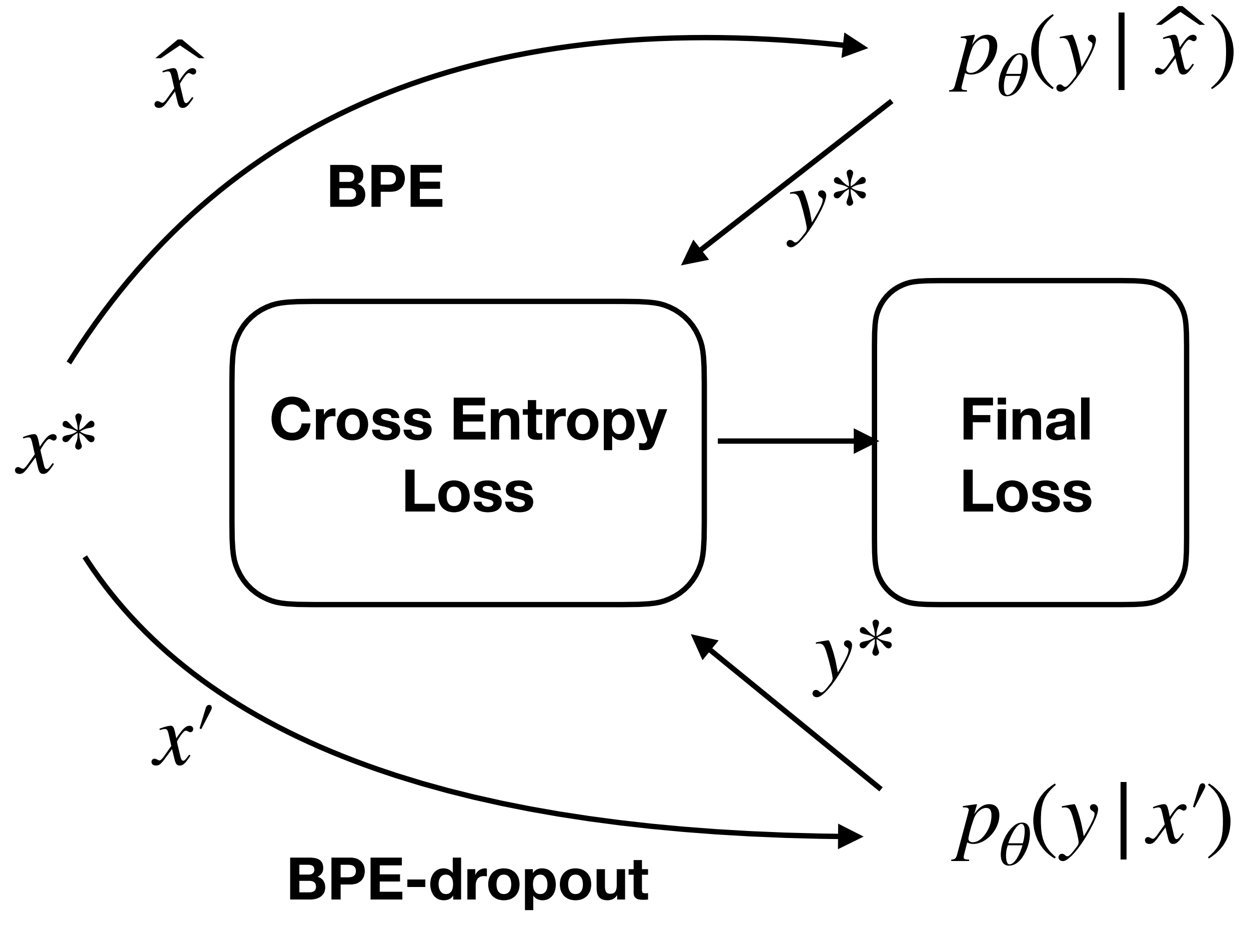
- Use both deterministically and probabilistically segmented inputs
- Enforce the prediction consistency between the two inputs

Multi-view Subword Regularization (MVR)



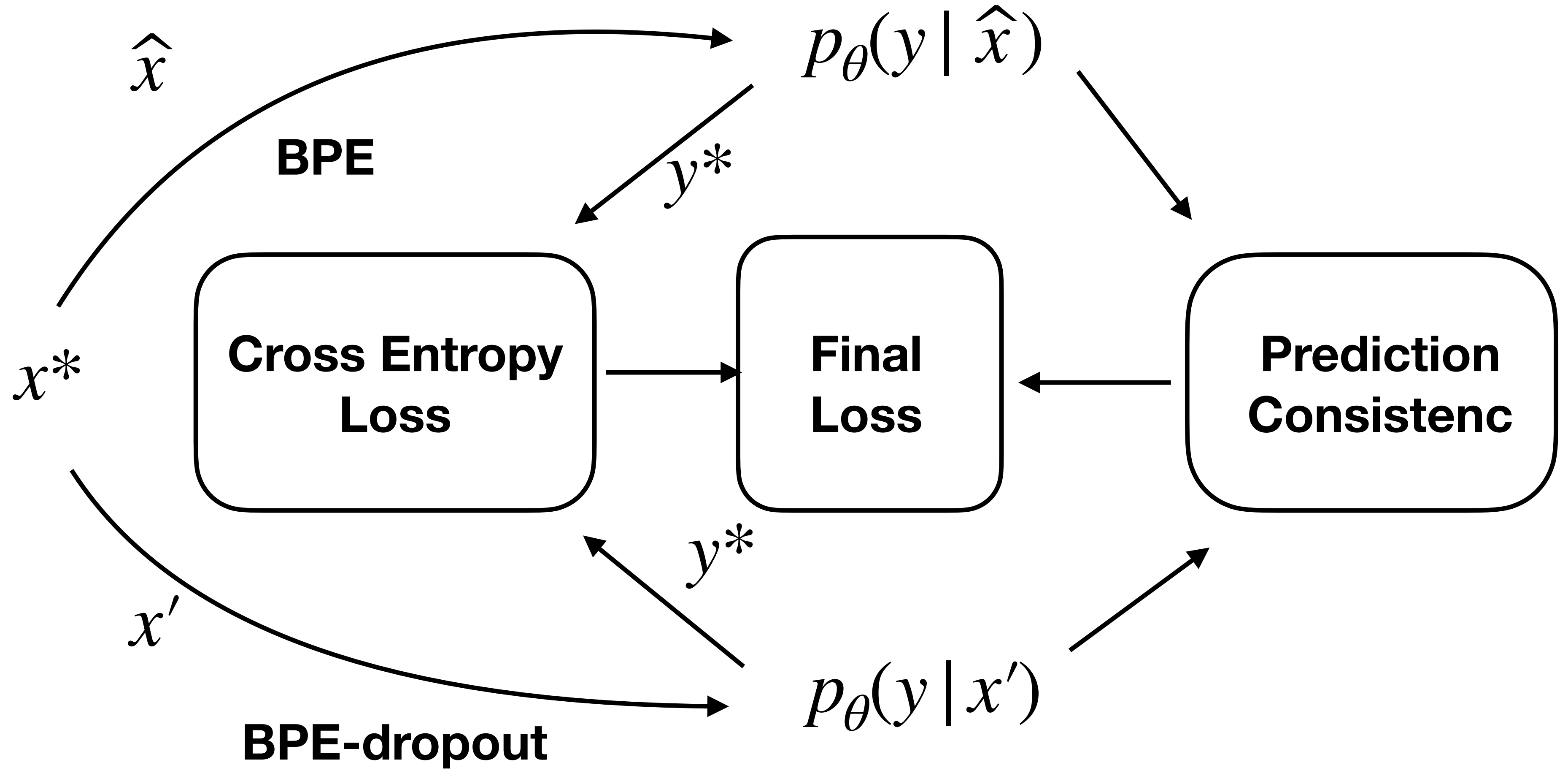
- Deterministic seg. CE: maximizes the benefit of pretraining

Multi-view Subword Regularization (MVR)



- Probabilistic seg. CE: allows the model see different segmentations

Multi-view Subword Regularization (MVR)

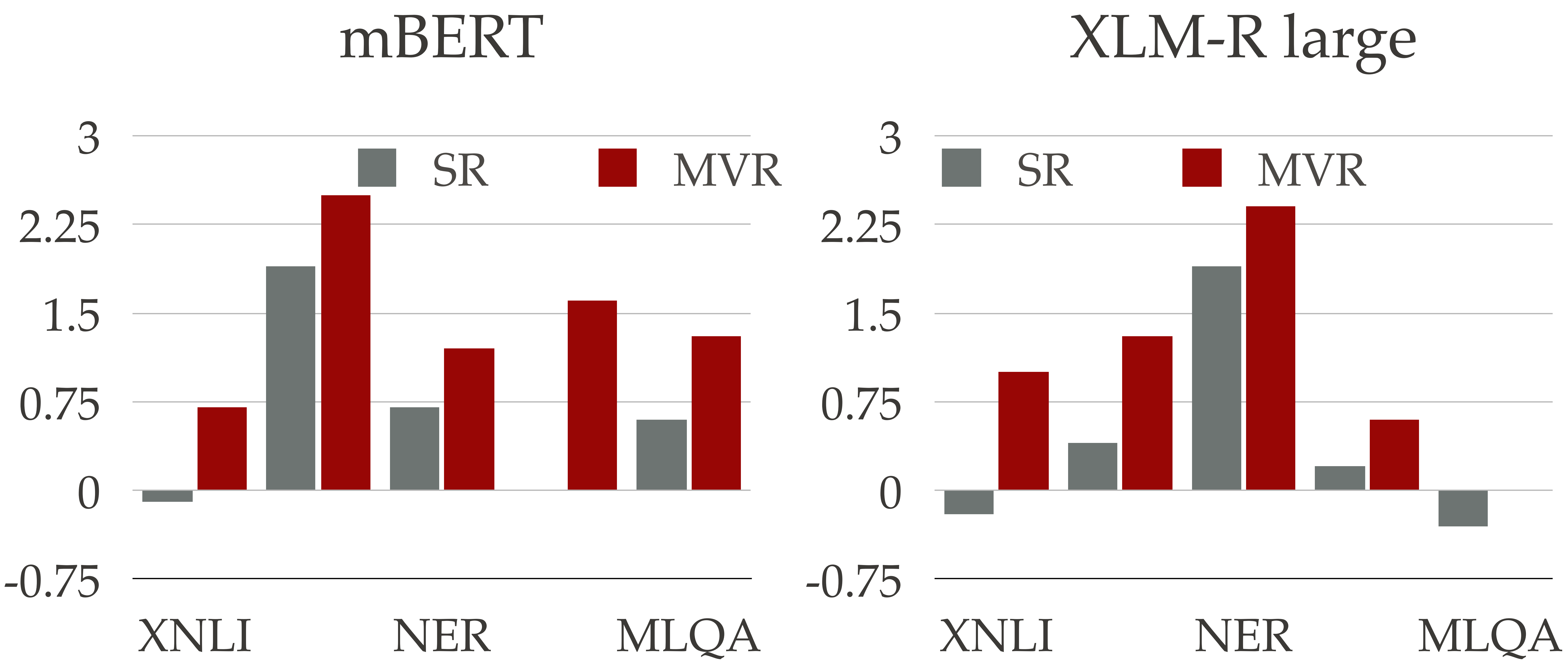


- Consistency loss: enforces the model to make consistent prediction, which improves the robustness to segmentation of multilingual data

Experiments

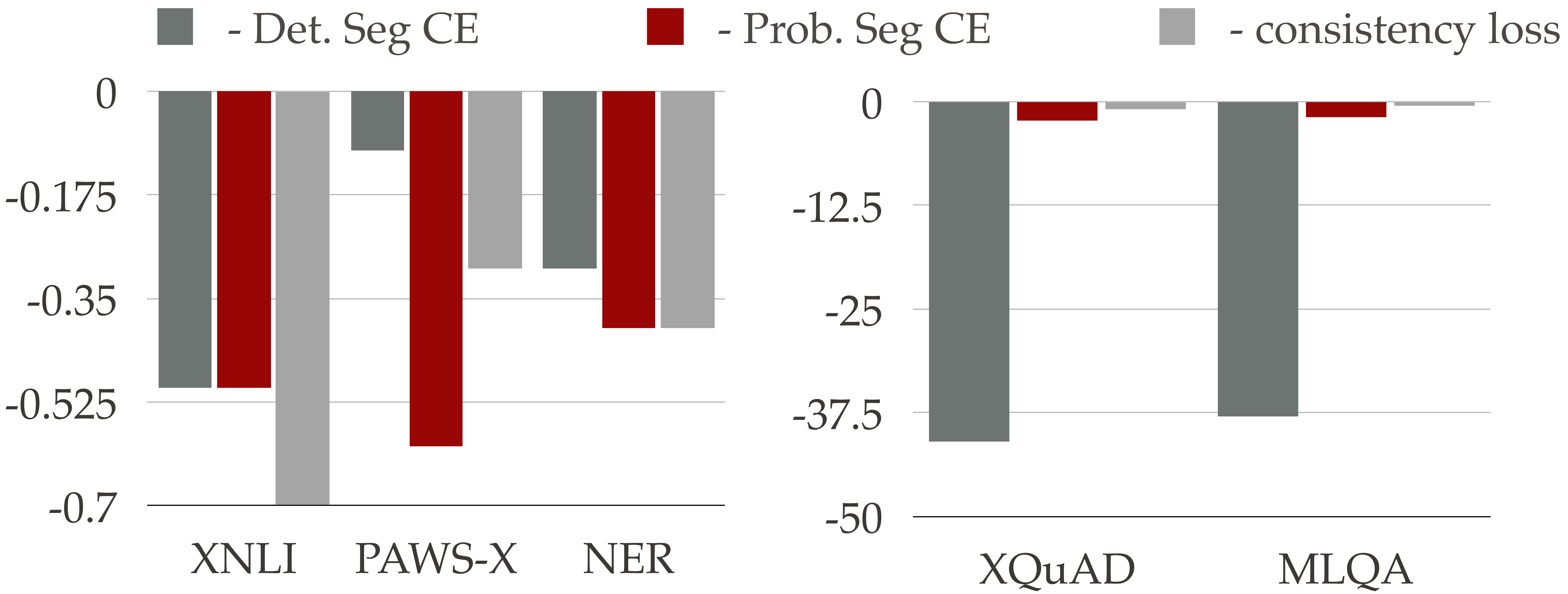
- XTREME tasks (Hu et. al. 2020)
  - Tagging: NER
  - Classification: XNLI, PAWS-X
  - QA: XQuAD, MLQA
- Model
  - mBERT
  - XLM-R base, large

Results



- Applying SR on English significantly improves other languages
- MVR consistently improves over SR

Ablations



- Removing any of the components hurts performance
- Det. Seg CE has large effect on QA probably because prob. seg clashes with span extraction

Latin vs. non-Latin script

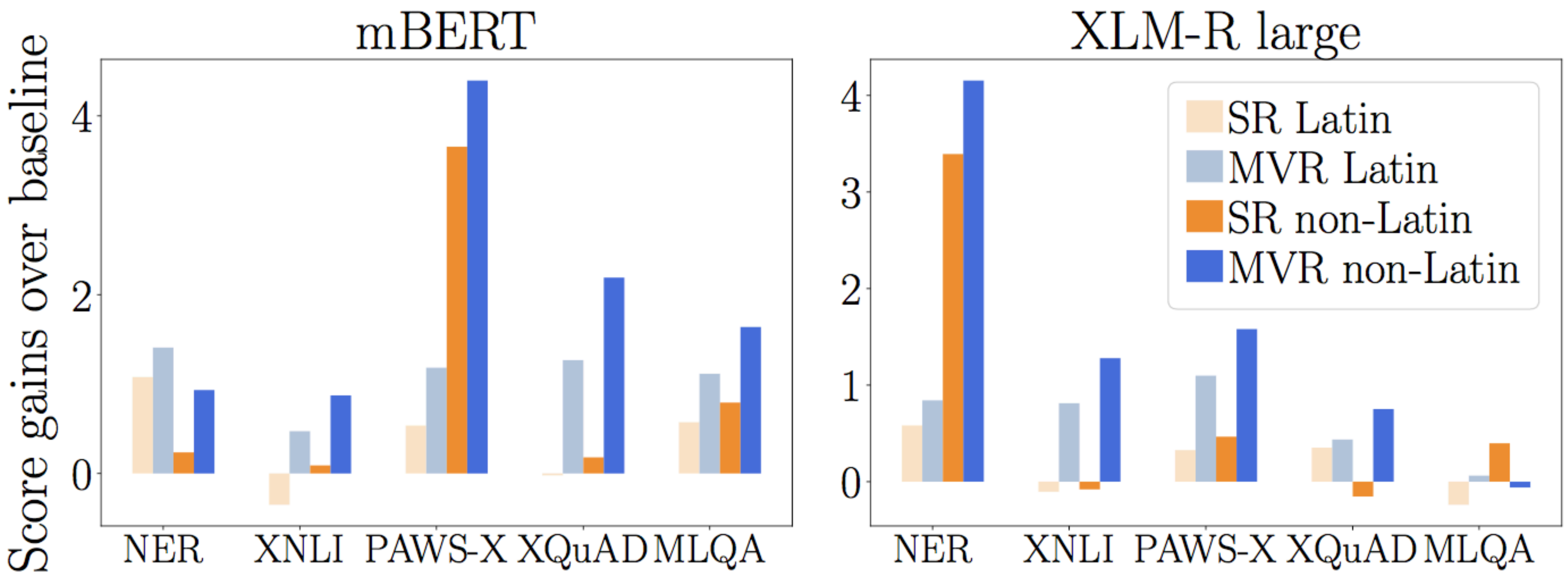


Figure. Improvements over baseline for Latin vs. non-Latin languages

- Both MVR and SR improve more for non-Latin languages

Effect on over-segmentation

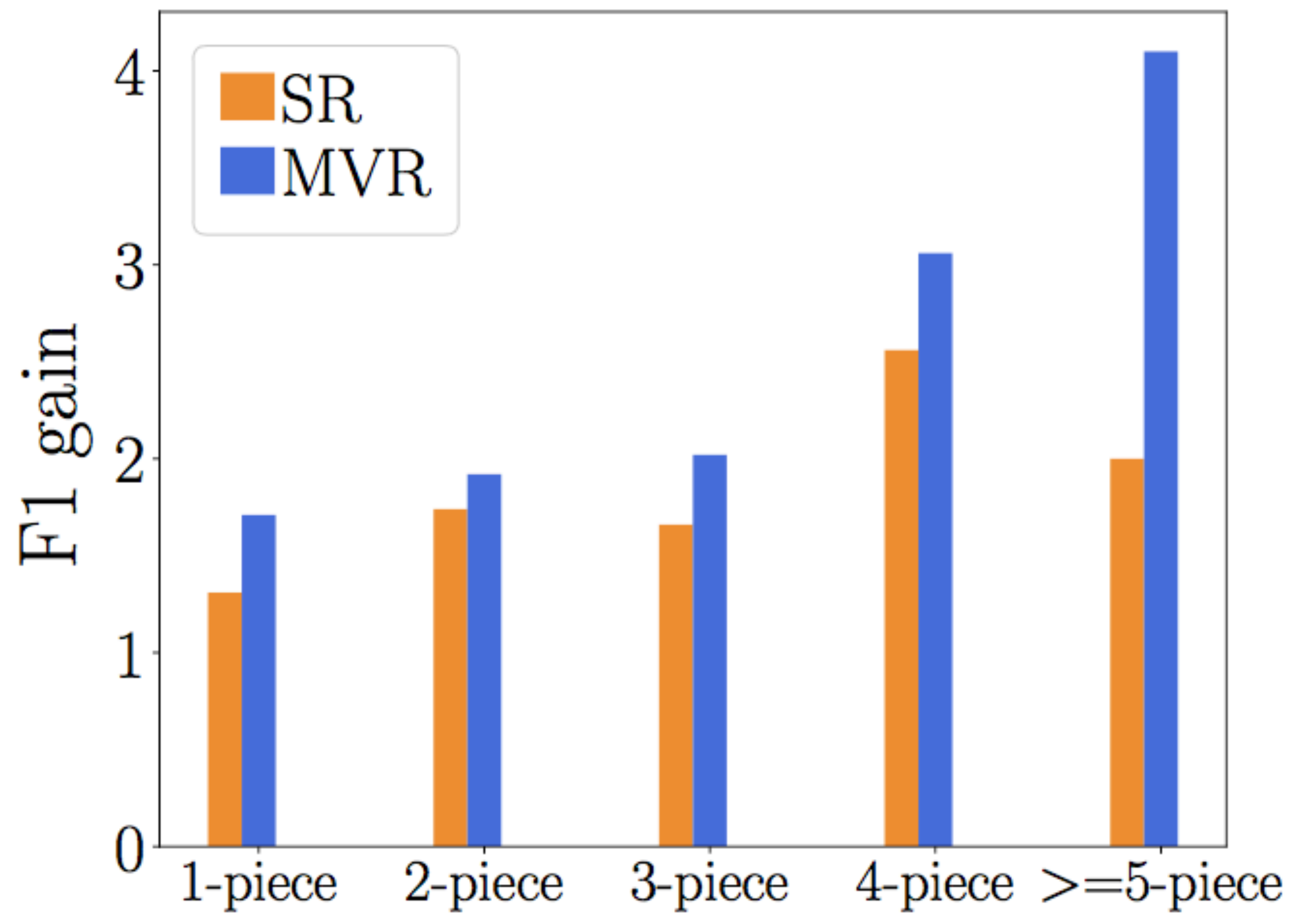


Figure. XLM-R large gains over NER baseline

- MVR tends to improve more for words segmented into large number of pieces

Effect of consistency loss

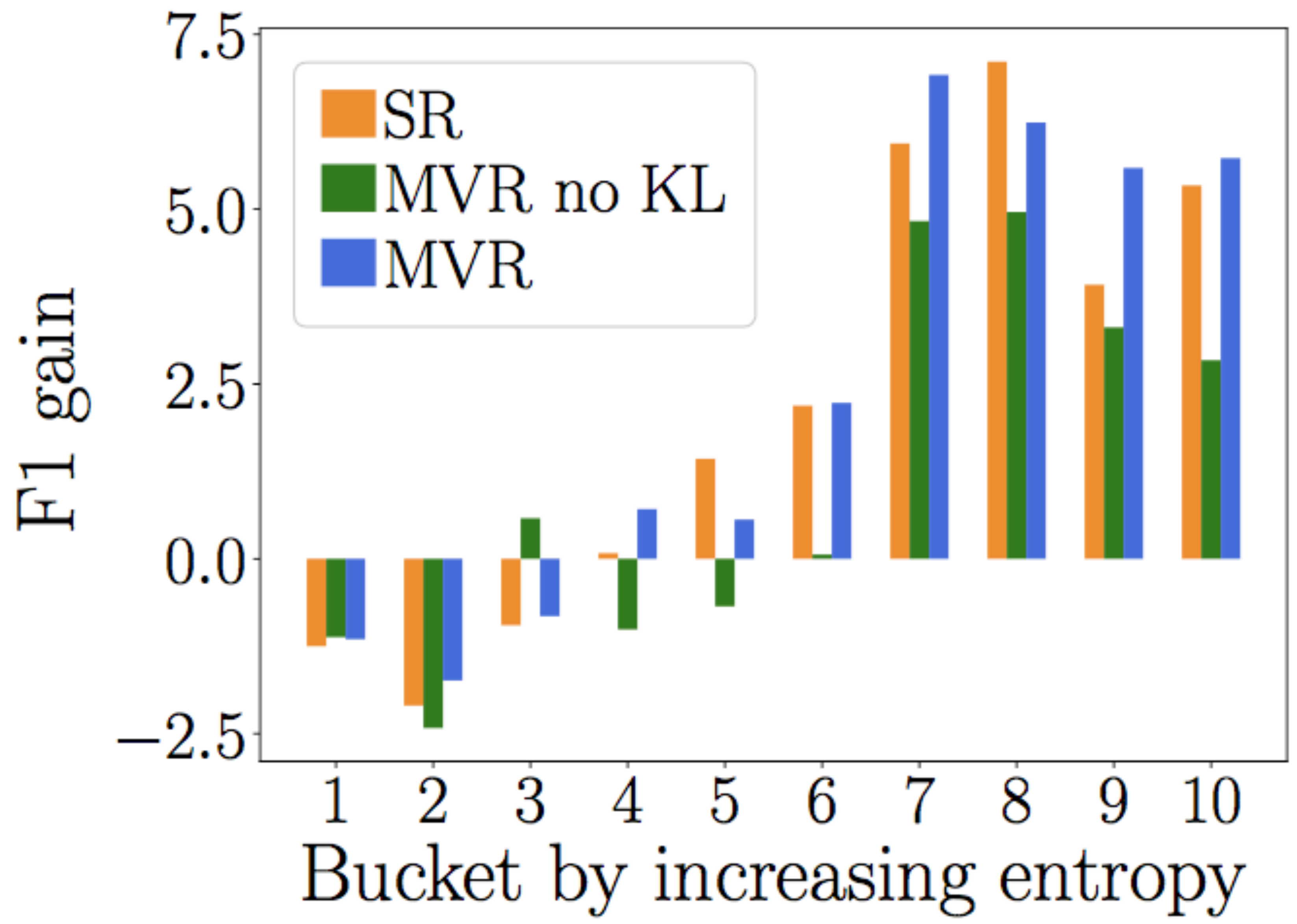


Figure. mBERT gains over NER baseline

- Consistency loss helps examples with higher entropy
- Label smoothing effect: calibrate the two predictions against each other

Effect of consistency loss

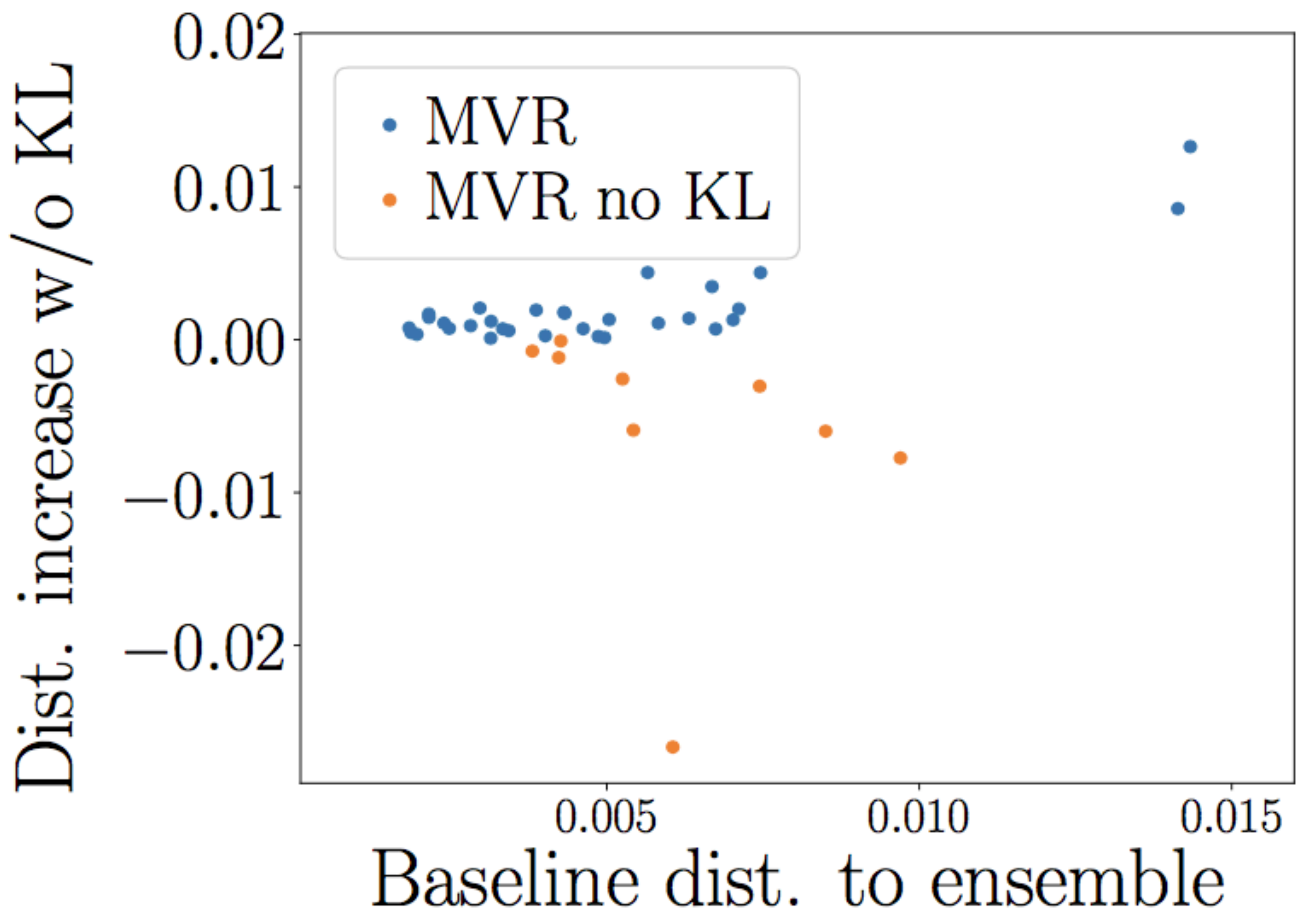


Figure. Full MVR is closer to ensemble distribution

- Languages colored by the method leading to closer distribution to the ensemble of baseline and SR models
- Ensemble effect: Consistency loss shifts model prediction closer to the ensemble

Effect on English

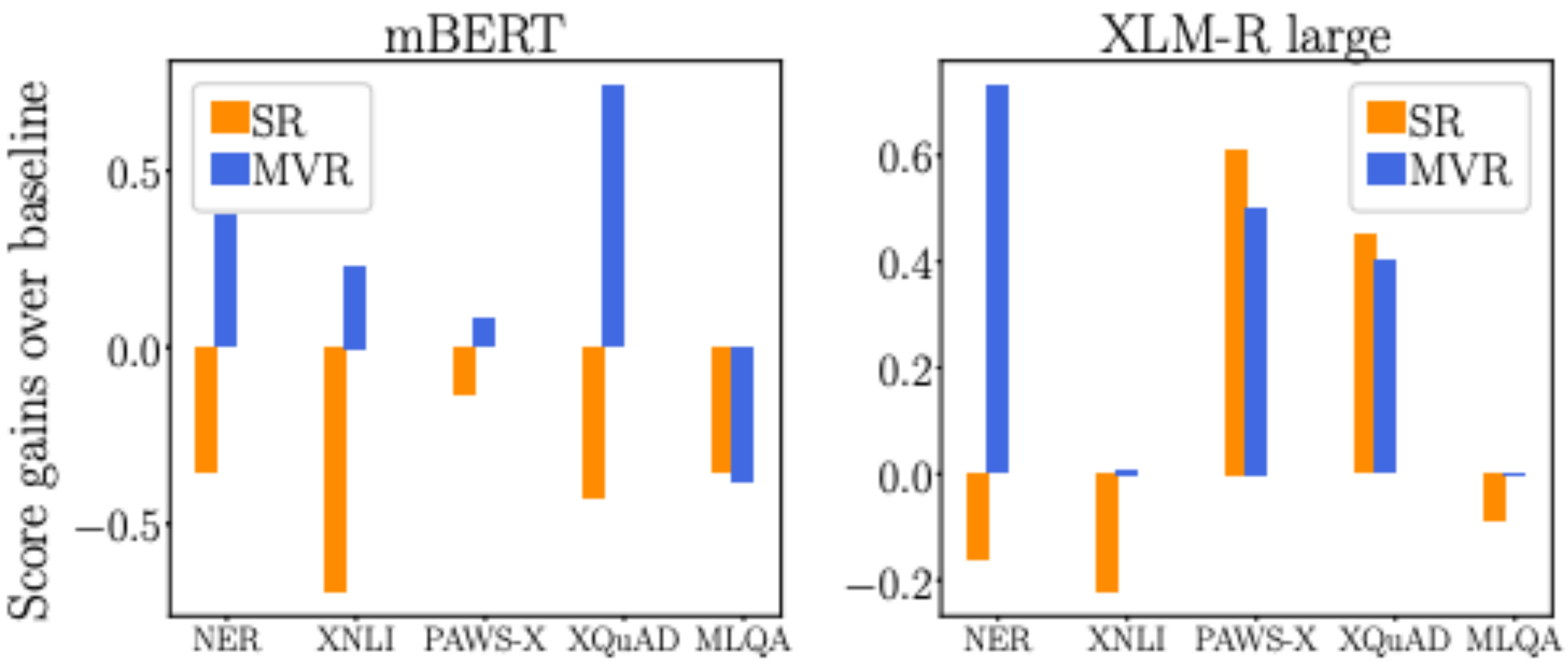


Figure. Gains of MVR and SR for English

- SR sometimes harm the performance of English, especially on XLM-R large
- MVR generally improves over the baseline and SR on English

Conclusion

- Deterministic word segmentation is **sub-optimal** for multilingual pretraiend models
- **Simple subword regularization** at fine-tuning can improve performance
- Multi-view Subword Regularization further brings **consistent improvements**

- **Code**: https://github.com/cindyxinyiwang/multiview-subword-regularization
- **Questions/comments**: xinyiw1@cs.cmu.edu