T5

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

Less Data, More ___?
Data Augmentation and Semi-Supervised Learning for Natural Language Processing

Diyi Yang, Georgia Tech
Ankur P. Parikh, Google Research
Colin Raffel, University of North Carolina, Chapel Hill

Data Augmentation

- Token-level augmentation
  - Change individual words
- Sentence-level augmentation
  - Change an entire sentence
- Adversarial augmentation:
  - Change the text to maximally fool the model
- Hidden space augmentation:
  - Change the representations inside the model

Semi-supervised learning

- Consistency regularization
  - Train the model to output consistent predictions after augmentation
- Entropy regularization
  - Train the model to output confident predictions
- Self-training
  - Train the model to predict its own outputs
- How to find unlabeled data?
  - Mine unstructured text corpora for task-specific data
- Leveraging the pre-training format
  - Pre-training on downstream data and framing tasks as cloze problems

Applications to Multilinguality

- What should we do when we have limited data in some languages?
- Multilingual Pre-training
  - Pre-train the model on a large multilingual corpus
- Back-Translation for Machine Translation
  - Generate additional data through paraphrasing
- Zero shot Translation
  - Translate between unseen language pairs
- Unsupervised Machine Translation
  - Translate without any paired data

Data augmentation

1. Token-level augmentation
   - Synonym replacement (e.g. back roads -> backward)
   - Random insertion, deletion, swapping
   - Word replacement via LM

     (문장의 뼈대를 유지한 채 특정 단어를 문맥에 기반해 예측된 단어로 교체)

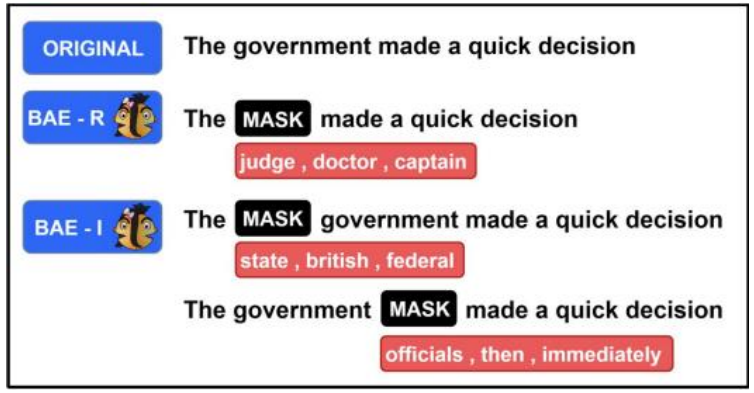
2. Sentence-level augmentation
   - Paraphrasing
   - Conditional generation (언어모델에 기반한 데이터 증강)

Data augmentation

3. Adversarial augmentation
   - Whitebox methods
     (해당 텍스트에서 가장 중요한 글자를 찾고 그 한 글자를 변경)
   - Blackbox methods
     (특정 단어를 마스킹 하거나 단어 사이에 마스크를 추가하여 언어 모델이 예측한 적대적 예시 생성)

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mood of optimism.
57% **World**

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a moo**P** of optimism.
95% **Sci/Tech**



4. Hidden space augmentation
   (숨겨진 표현을 조작합니다.)

   - 노이즈를 추가하거나 다른 data points로 보강합니다.

- 모든 작업에 대해 단일 증강이 최고의 성능을 내진 않습니다.

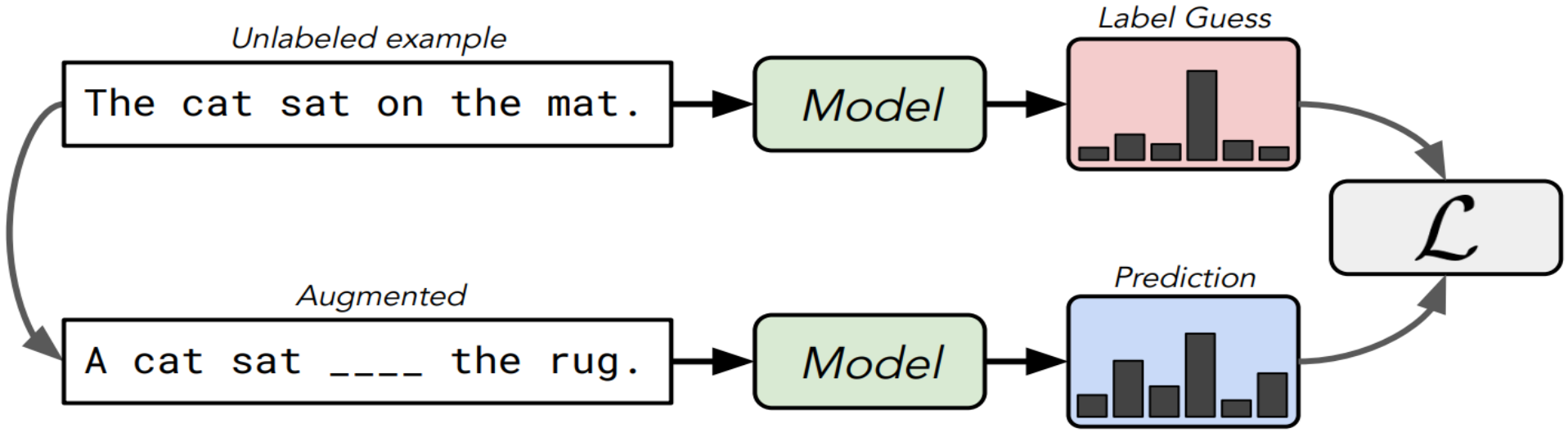
- 증강이 언제나 성능을 향상시키는 것은 아니며 때때로 성능을 손상시키기도 합니다.

- Token-level 증강은 일반적으로 지도학습, 특히 제한된 수의 라벨링 된 데이터에서 잘 작동합니다.

Semi-supervised learning

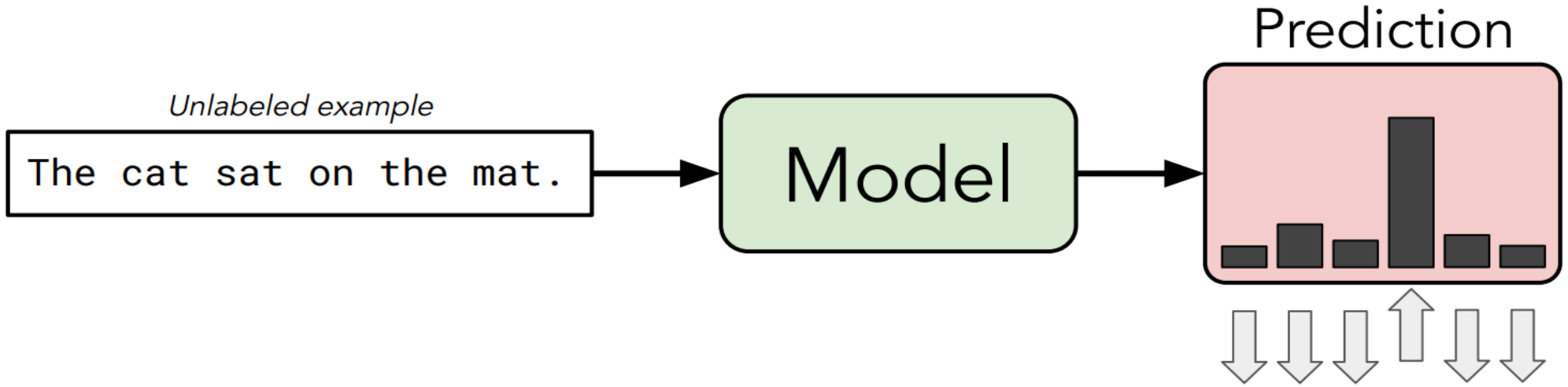
Consistency regularization

증강 후에도 일관된 예측을 출력하도록 모델을 학습시킵니다.



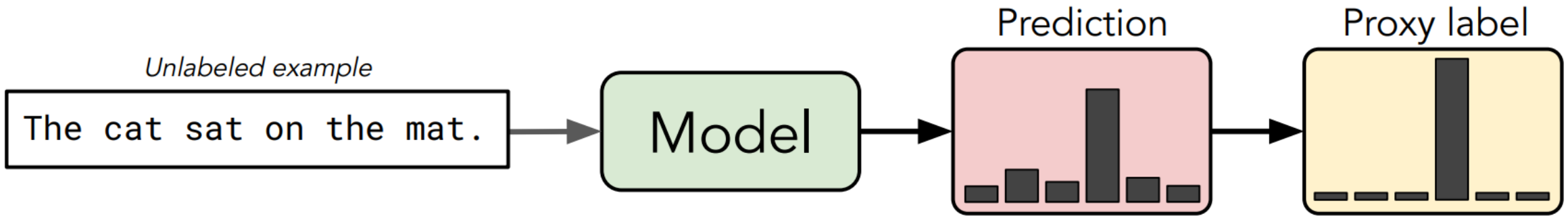
Entropy regularization

신뢰할 수 있는 예측을 출력하도록 모델을 학습시킵니다.



Self-training

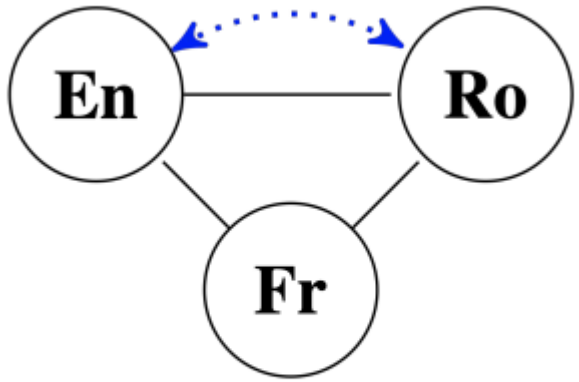
스스로의 출력을 예측하도록 모델을 교육합니다.



Applications to Multilinguality

- Back Translation
- Zero shot Translation
- Unsupervised Machine Translation

- Back-Translation



Supervised (Multilingual) Translation
[Johnson et al. 2016,
Firat et al. 2016]

Solid lines indicate presence of parallel data

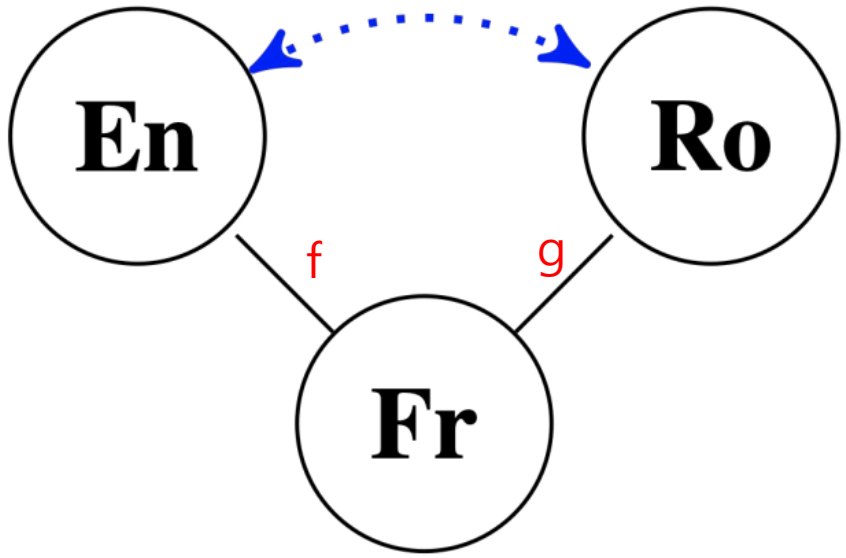
Paraphrasing을 통해 추가 데이터를 생성합니다.

순방향(Ro -> En) 모델의 성능이 향상될 가능성이 높으므로
순방향 모델로 합성 데이터를 생성하면
역방향(En -> Ro) 모델에 대한 고품질의 훈련 데이터를 생성할 수 있습니다.

따라서 Back-Translation은 데이터가 많은 언어에서
적은 언어로 번역하는 데 큰 도움이 될 수 있습니다.

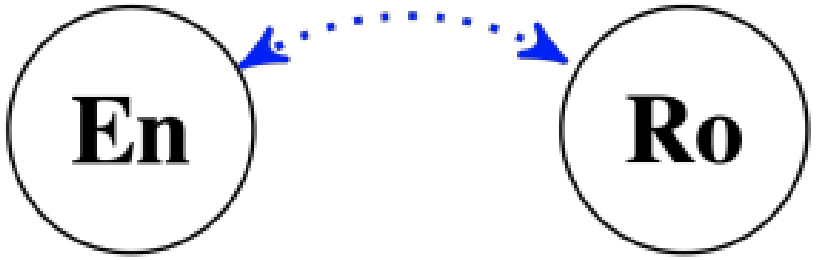
Synthetic Data Generation (Distillation) for Zero-Shot Translation [Chen et al. 2017]

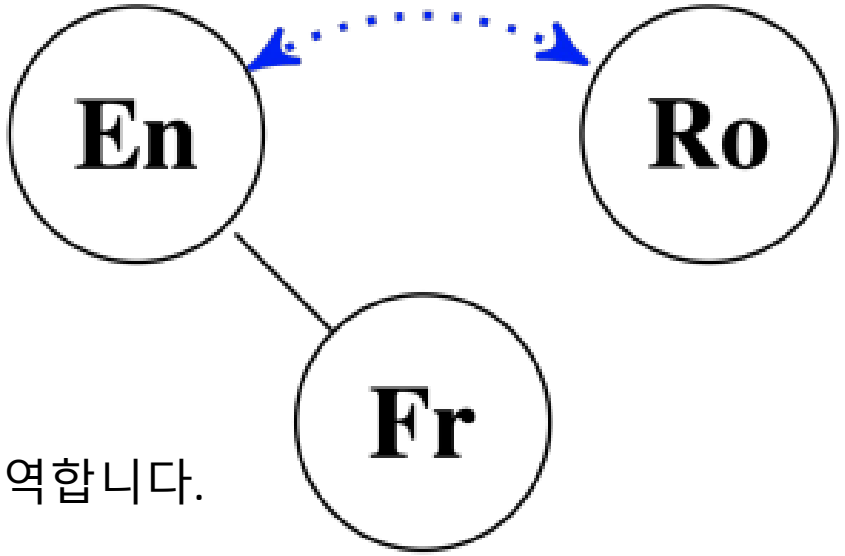
처음 보는 언어 쌍을 번역합니다.



- Train supervised (*En*, *Fr*) model f and (*Fr*, *Ro*) model g
- Use g to label (*En*, *Fr*) data to generate synthetic (*En*, *Ro*) data
- Train (*Er*, *Ro*) model

Unsupervised Machine Translation





쌍으로 구성된 데이터 없이 번역합니다.

Unsupervised translation [Ravi and Knight 2011, Lample et al. 2018, Artexe et al. 2018]

Multilingual Unsupervised Translation [Siddhant et al. 2020, Garcia et al. 2020, Li et al. 2020, Wang et al. 2021, Garcia et al. 2021]

Solid lines indicate presence of parallel data

Step 1: Train a pretrained language model based on monolingual data in the source and target languages (with span denoising)

Step 2: Use online back-translation:

감사합니다