DETECTING HALLUCINATED CONTENT IN CONDITIONAL NEURAL SEQUENCE GENERATION

Carnegie Mellon University School of Computer Science
Language Technologies Institute

Chunting Zhou[1], Graham Neubig[1], Jiatao Gu[2], Mona Diab[2], Paco Guzman[2], Luke Zettlemoyer[2], Marjan Ghazvininejad[2]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Facebook AI Research

facebook Artificial Intelligence

Abstract

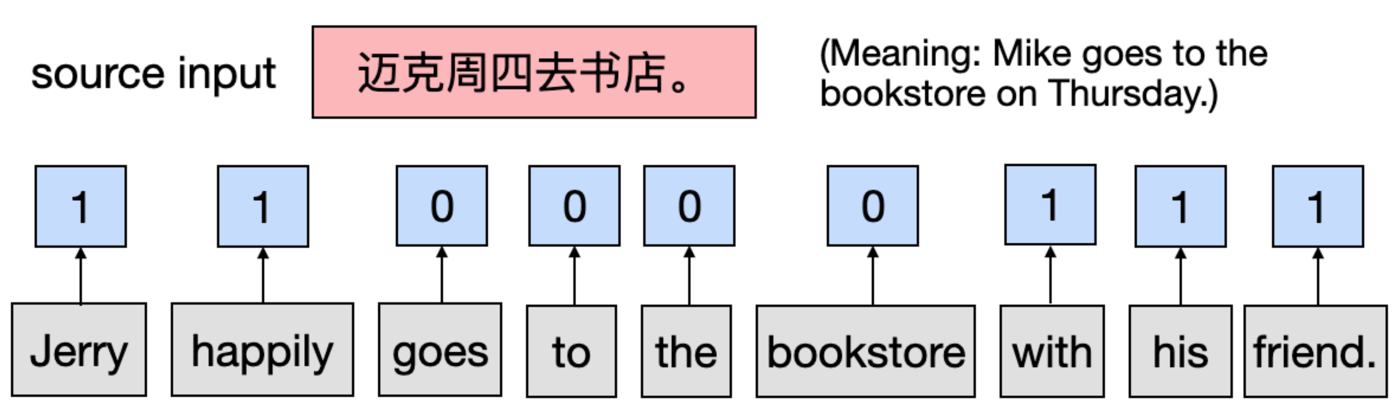
1. Conditional neural sequence generation systems can hallucinate new content not supported by the source inputs.
2. We develop an unsupervised method with pre-trained language language models to ***detect hallucinated tokens*** in the machine generation.
3. We propose a token-level truncated loss based on the outputs of our hallucination detection system to ***improve noisy training*** where training data contains hallucinated noise.

Code available at:
https://github.com/violet-zct/fairseq-detect-hallucination

Hallucination: **fluent text output but not supported by the input.**

— neural machine translation in out-of-domain or low-resource setting
— abstract summarization (Maynez et al., 2020)
— extrinsic (additional content) v.s. intrinsic (synthesized content) hallucinations

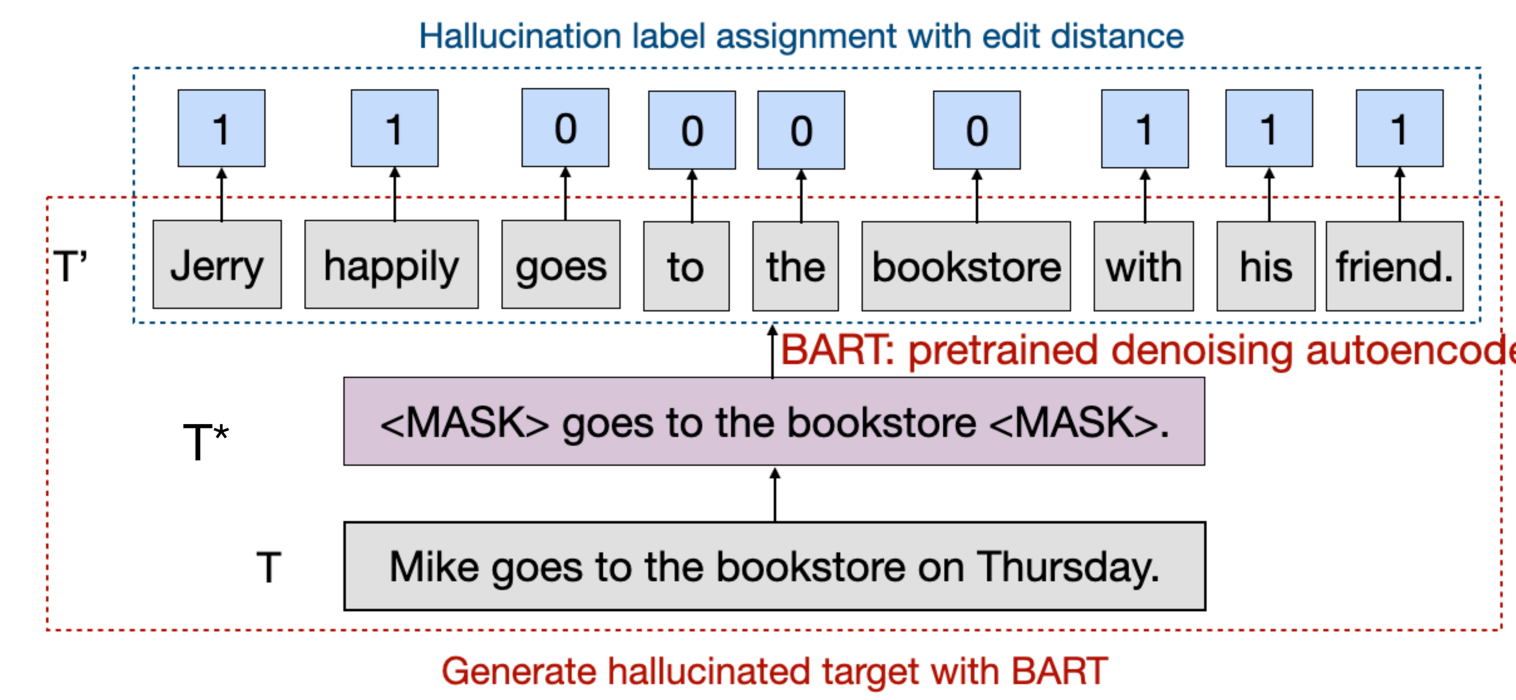
TOKEN-LEVEL HALLUCINATION PREDICTION: AN EXAMPLE IN MT



A TWO-STAGE HALLUCINATION DETECTION MODEL

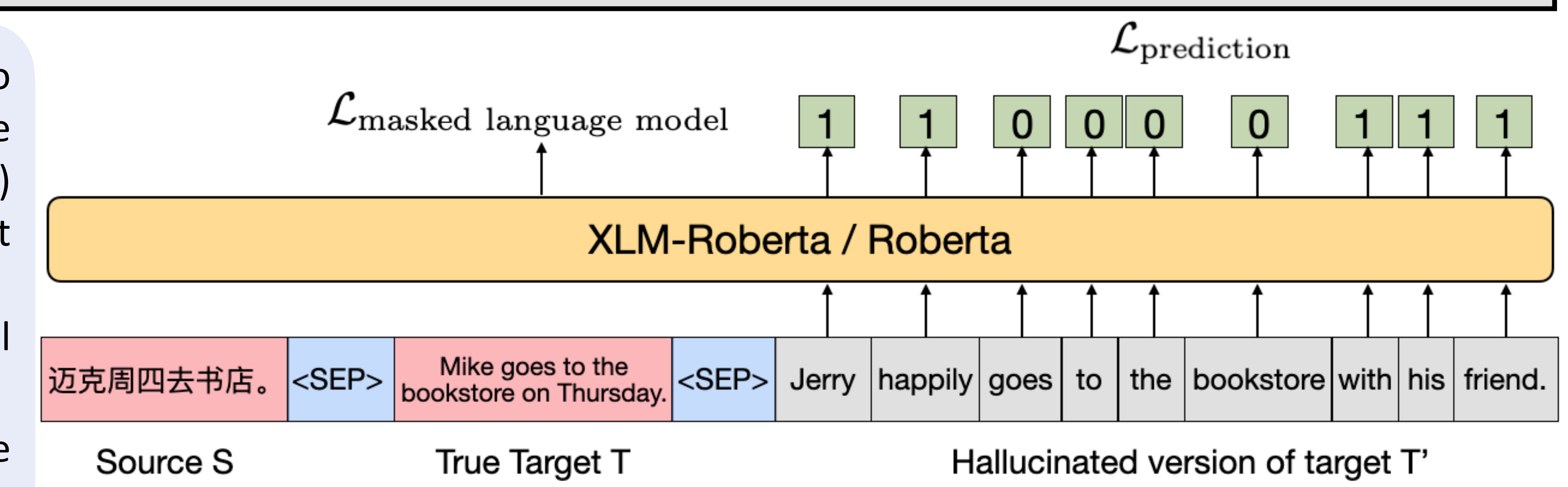
Step1: Synthetic Labeled Data Creation

1. Given the target sequence T in the bi-text training set, a hallucinated version of it T* is created by first corrupting T with noise functions.
2. T* is fed into the pre-trained denoising autoencoder BART to generate a new sentence T'.
3. Finally, each token in T' is assigned the pseudo hallucination label by computing the edit-distance between T' and T.



Step2: Fine-tuning a Pretrained LM on the Synthetic Data

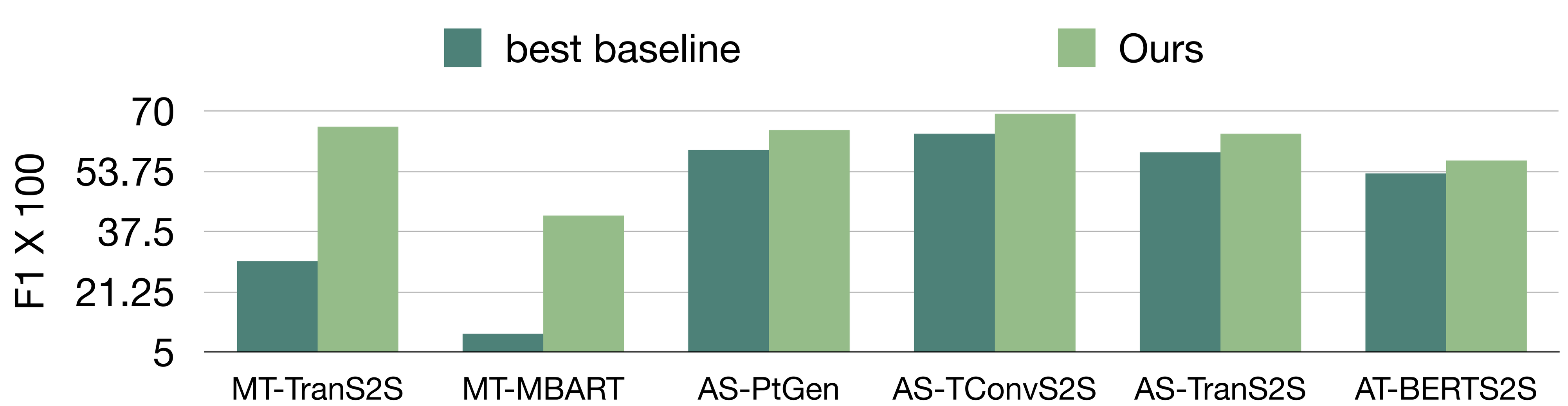
- Given the synthetic data (T' and its pseudo labels), we fine-tune a pre-trained language model XLM-Roberta (for e.g. machine translation) or Roberta (for e.g. summarization) to predict hallucination labels for the target side.
- We also add a multi-task masked language model objective on the true targets.
- At test time, we only concat S and machine generation as the input, and generalizes well.



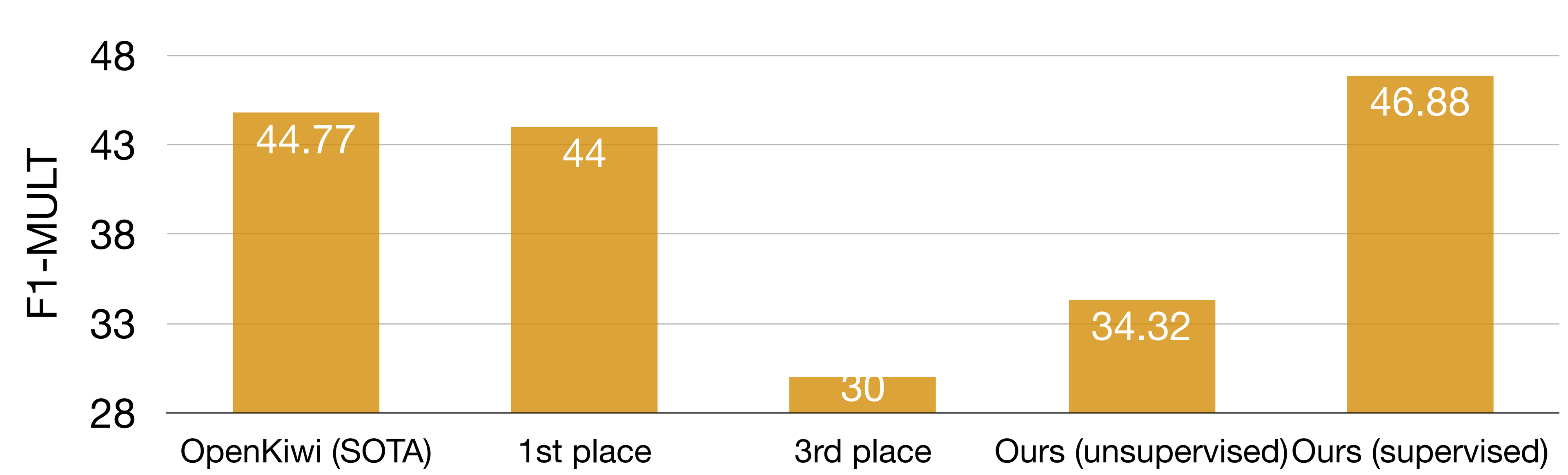
Training objective: $\mathcal{L} = \mathcal{L}_{\text{pred}} + \alpha \cdot \mathcal{L}_{\text{mlm}}$

EVALUATION ON TOKEN-LEVEL HALLUCINATION DETECTION

- We evaluate on 4 abstract summarization (AS) test sets (XSum, Maynez et al., 2020) and 2 machine translation (MT) test sets that we created. We proposed three strong baselines for this new task.
- We show the F1 of hallucination labels. Ours outperforms baselines significantly, especially on MT.



EVALUATION ON WORD-LEVEL QUALITY ESTIMATION (WMT18)



LEVERAGING HALLUCINATION LABELS IN NOISY TRAINING

Case I: Improving Self Training in Machine Translation

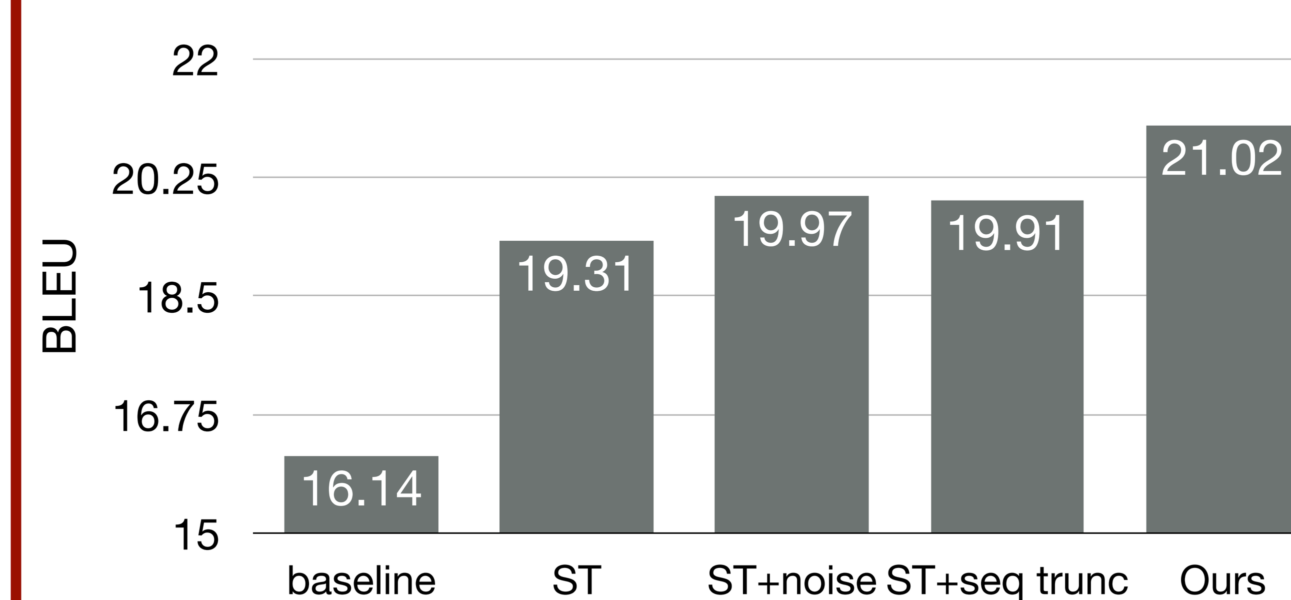


Fig. BLEU scores, ours—self-training with token-level truncation loss

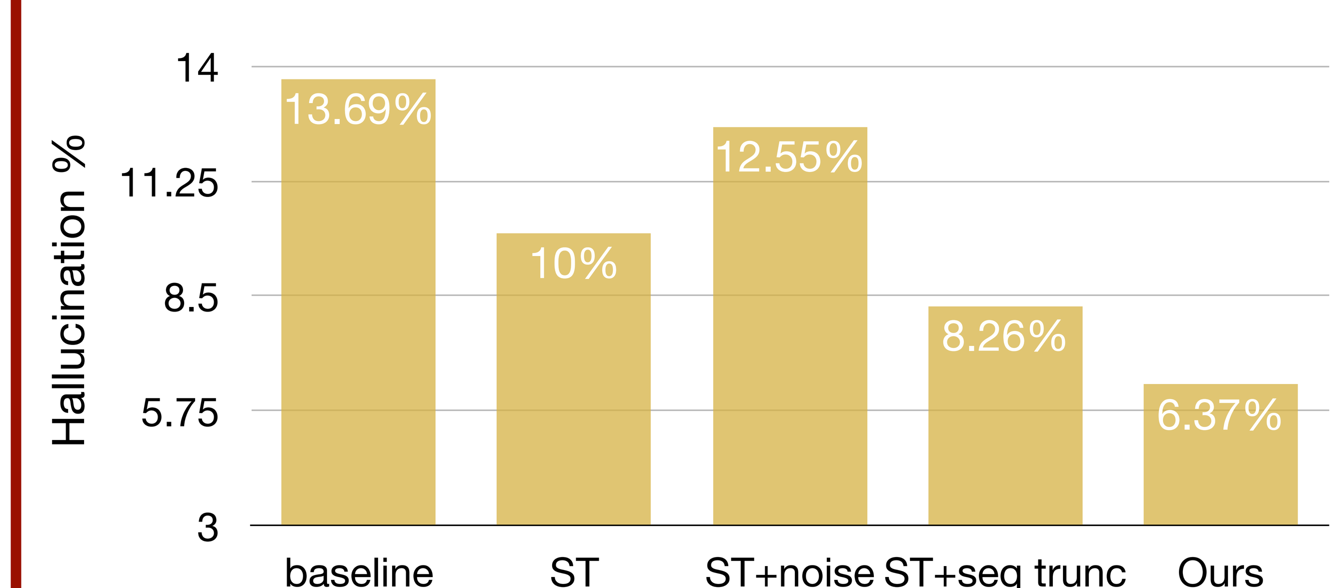


Fig. Percentage (%) of hallucination tokens in the machine translations

- Token-level hallucination labels are fine-grained signals.
  - We use a fine-grained loss for noisy training: excluding predicted hallucinated tokens $H(y)$ in the noisy target y: $\ell(y|x;\theta) = \sum_{i \leq N; y_i \notin H(y)} \log p(y_i | y_{<i}, x; \theta)$
- Reduce adverse effects of noisy training instances by *maximally using the clean part*
  - self-training with weak teacher model can produce noisy pseudo targets
  - training data of low-resource language pairs are often in low-quality

Case II: Improving Corpus Filtering for Low-Resource MT