Motivation: Context-Entity Mismatch Issue in Data Augmentations

1. Complex Named Entity Recognition (NER) faces the issues of syntactically ambiguous and linguistically complex Named Entities (NEs), low context sentences, and emerging NEs which PLMs might not have seen in their pre-training data. These problems lead current NER data augmentation frameworks to generate incoherent augmentations with context-entity mismatch (where NEs do not match the surrounding context). They also impede a model from learning effective NE patterns, thereby leading to sub-optimal performance.

2. We propose ACLM (Attention-Guided Conditional Language Model fine-tuning), a novel framework that optimizes BART on a novel text-reconstruction task based on selective masking. The resulting augmentations are coherent and diverse and outperform all our baselines.

Proposed Approach: Selective-Denoising with Attention Maps

4 step sentence corruption for sequence-to-sequence denoising-based fine-tuning:

1. Keyword Selection using attention maps from fine-tuned transformer encoder to preserve valuable hints. The top-3% of tokens which are not NEs and have the highest attention provide explicit label supervision.

2. Selective Masking by masking all other tokens except of the selected keywords in Step 1. The top-p% of tokens which are not NEs and have the highest attention are masked.

3. Labeled Sequence Linearization: Dynamic Masking by randomly masking some keywords from Step 1 in each iteration to promote length diversity.


Additionally, we propose mixer that mixes 2 semantically similar sentences before step 4 to promote diversity.

Results: ACLM outperforms all baselines Qualitatively and Quantitatively.

Table 1: Monolingual and Cross-lingual Results: Improvement of 1% - 22%

Table 2: Multi-lingual Results: Improvement of 1% - 22%

Table 3: Results on 4 other domains

Table 4: Qualitative Results

ACLM: A Selective-Denoising based Generative Data Augmentation Approach for Low-Resource Complex NER

Sreyan Ghosh, Utkarsh Tyagi, Manan Suri, Sonal Kumar, S. Ramaneswaran, Dinesh Manocha

University of Maryland, College Park, USA, NSUT Delhi, India, NVIDIA, Bangalore, India

ACLM: A Selective-Denoising based Generative Data Augmentation Approach for Low-Resource Complex NER

Lorem ipsum