Lexical Substitution as Causal Language Modeling

Ning Shi, Bradley Hauer, Grzegorz Kondrak
{ning.shi,bmhauer,gkondrak}@ualberta.ca

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Alberta Machine Intelligence Institute (Amii).

Introduction

Lexical **S**ubstitution **T**ask (LST) is to identify suitable replacements for a target word while preserving the contextual meaning of the sentence.

LST(S, w_x) = y
Sentence (S) = "Let me <u>begin</u> again."; Target Word (w_x) = "begin"
Substitutes (y) = ["start", "commence", "open", ...]

Existing methods involve contextualized representations and sequence-to-sequence generation; however, several **limitations** remain.

We provide the first single-step, end-to-end generative solution for LST that can also address existing limitations.

Methodology

We solve LST by reducing the problem to **W**ord **P**rediction (WP), which we solve via **C**ausal **L**anguage **M**odelling (CLM).

Lexical **Sub**stitution, **LexSub**(S, w_x, w_y) := *"the word w_x can be replaced by the word w_y in the sentence S without altering its meaning"*

LexSub("Let me <u>begin</u> again.", "begin", "start") = True

WP(S, w) := *"the word w has the same meaning as the masked word in the sentence S"*

Task Reduction from **LexSub** to **WP**:

$$LexSub(S, w_x, w_y) \Leftrightarrow WP(S, w_x) \wedge WP(S, w_y)$$

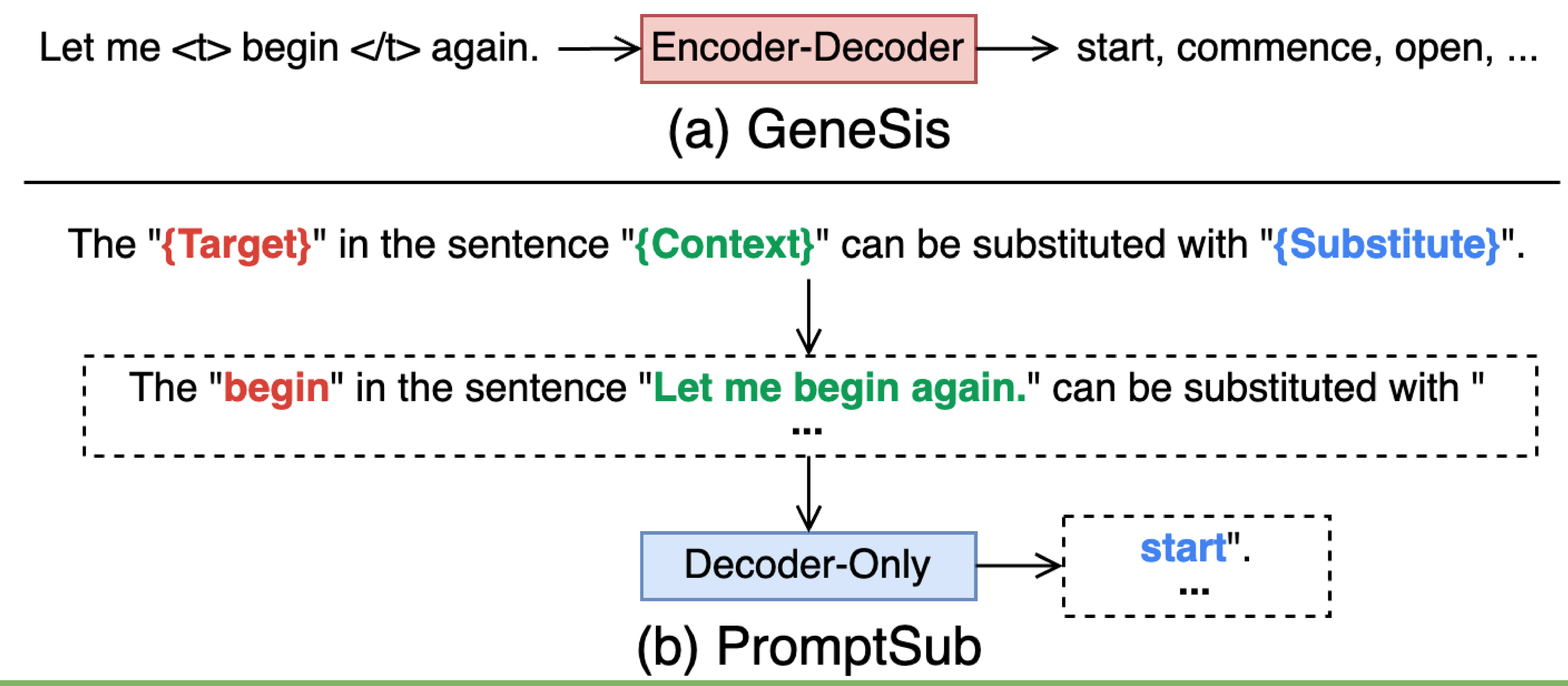
Limitations of Prior Work

Semantic Change:
Predicted substitutes may align well with the context but significantly *change the original meaning* of the sentence.

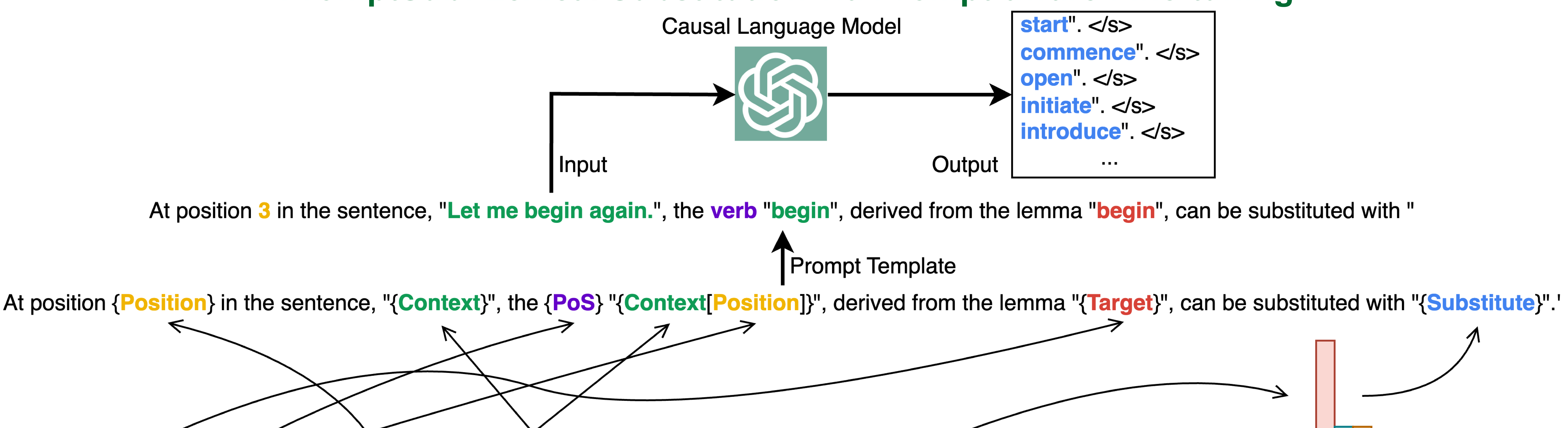
"Let me <u>carry on</u> again." vs. "Let me <u>originate</u> again."

Pipeline Approach:
(1) Manually defined heuristics and tuned thresholds
(2) Extensive pre-processing and post-processing steps
(3) Dependence on expert knowledge and external resources

Misalignment: Pre-training (language modeling) vs. Fine-tuning (LST)
Model Architecture (e.g., GeneSis):



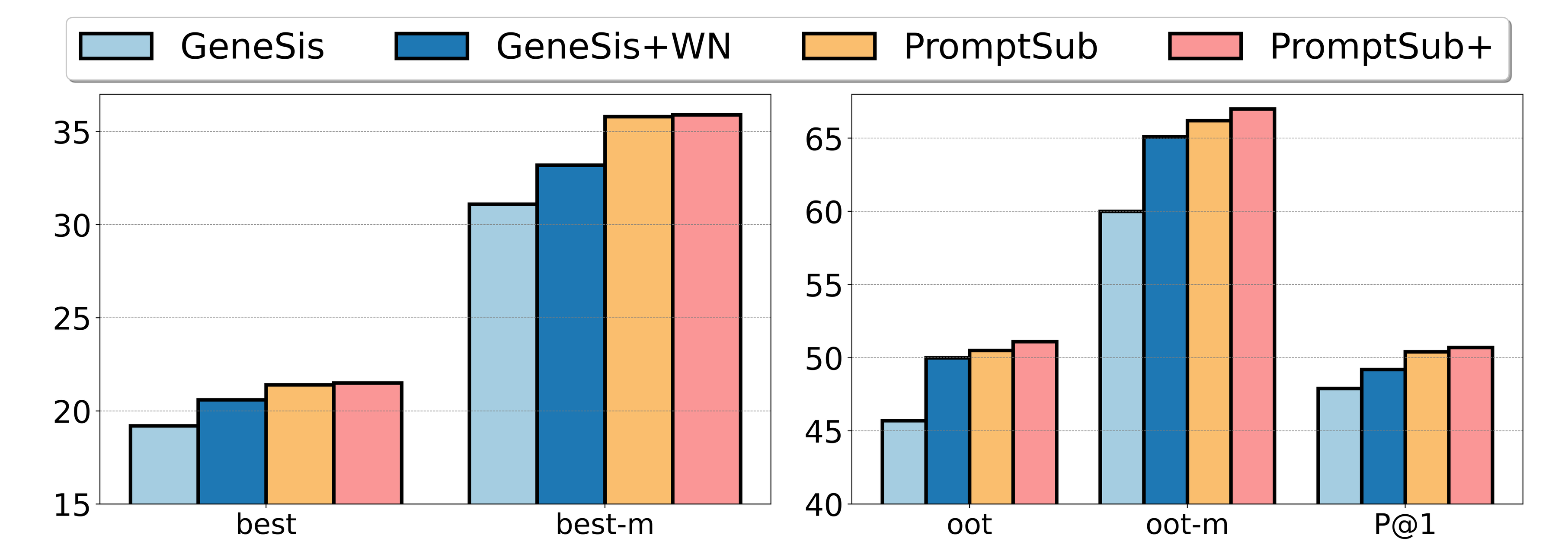
PromptSub: Lexical Substitution via Prompt-aware Fine-tuning



Target	PoS	Position	Context	Substitute::Frequency
begin	verb	3	Let me begin again.	start::6, commence::2, open::2, initiate::1, introduce::1, ...

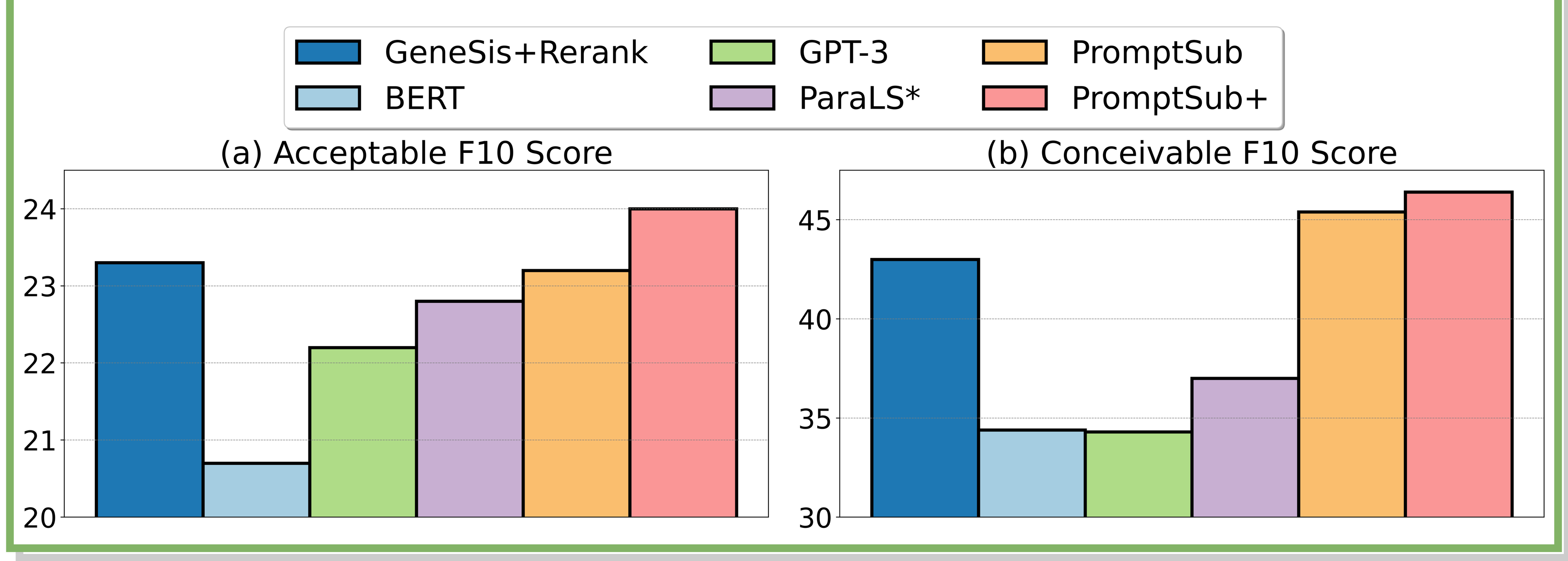
LS07 Results

PromptSub+ augments the training set by incorporating the dev set.



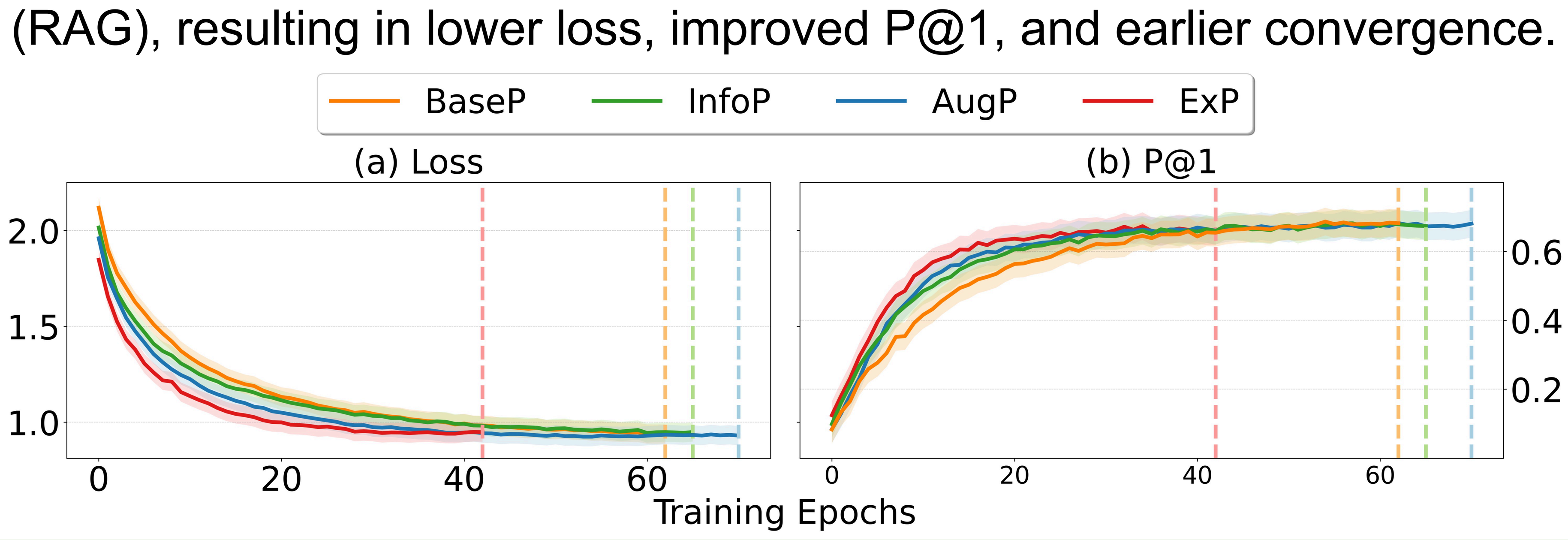
LS21 Results

We reproduced GeneSis+Rerank, which incorporates post-processing steps to refine the results.



Analysis

ExP retrieves WordNet synsets for retrieval-augmented generation (RAG), resulting in lower loss, improved P@1, and earlier convergence.



Conclusion

- An innovative and successful attempt to apply CLM to LST through a formally defined task reduction.
- A new state of the art on the LS21 benchmark by a large margin.
- Scalability via data resources, model capacity, and RAG.

github.com/ShiningLab/PromptSub

@*SEM 2024