Lexical Substitution as Causal Language Modeling

Ning Shi, Bradley Hauer, Greg Kondrak

Alberta Machine Intelligence Institute (Amii)
Department of Computing Science
University of Alberta, Edmonton, Canada

Lexical Substitution Task (LST)

LST is to identify suitable replacements for a target word while preserving the contextual meaning of the sentence.

LST(S, w_x) = y, for example:

Sentence (S) = "Let me begin again."

Target Word (w_x) = "begin"

Substitutes (y) = ["start", "commence", "open", ...]

Limitations of Prior Work

- Predicted substitutes may align with the context **BUT** change the original meaning of the sentence. Consider **M**asked **L**anguage **M**odeling (MLM):

Input S_0: "Let me begin again."

"Let me [MASK] again."

Output S_1: "Let me start again."
WordNet: (Verb) take the first step or steps in carrying out an action.

VS

Output S_2: "Let me originate again."
WordNet: (Verb) bring into being.

- Pipeline approaches, depending on defined **heuristics**, tuned **thresholds**, extensive **post-processing** steps, and external **resources**.
- A **GAP** between pre-training (language modeling) and fine-tuning (LST).

WordNet 3.1 & Open English WordNet.

Our Contributions

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

- An innovative and successful attempt to apply **C**ausal **L**anguage **M**odeling (CLM) to LST through a formally defined **task reduction**.
- A new overall **state-of-the-art** result.
- **Scalability** via data resources, model capacity, and retrieval-augmented generation (RAG).

Task Definition

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 begin again.", "begin", "start") = TRUE

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

WP("Let me [~~begin~~] again.", "start") = TRUE

Task Reduction

A **P-to-Q** reduction solves an instance of a problem **P** by combining the solutions of one or more instances of **Q**.

A **mutual** reductions of two problems to one another demonstrate their **equivalence**.

Task Reduction from LexSub to WP:

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

(Hauer and Kondrak, 2023)

Method – PromptSub
Lexical Substitution via Prompt-aware Fine-tuning

InfoPrompt incorporate three additional attributes of the target word:

- Lemma form (**Target**)
- Part of Speech tag (**PoS**)
- Position in the **Context** (**Position**)

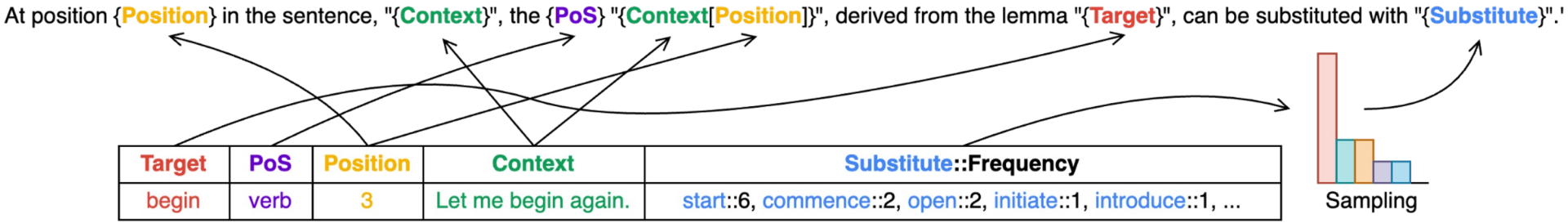
Exclusively from the task data thus **NO** reliance on external resources.

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

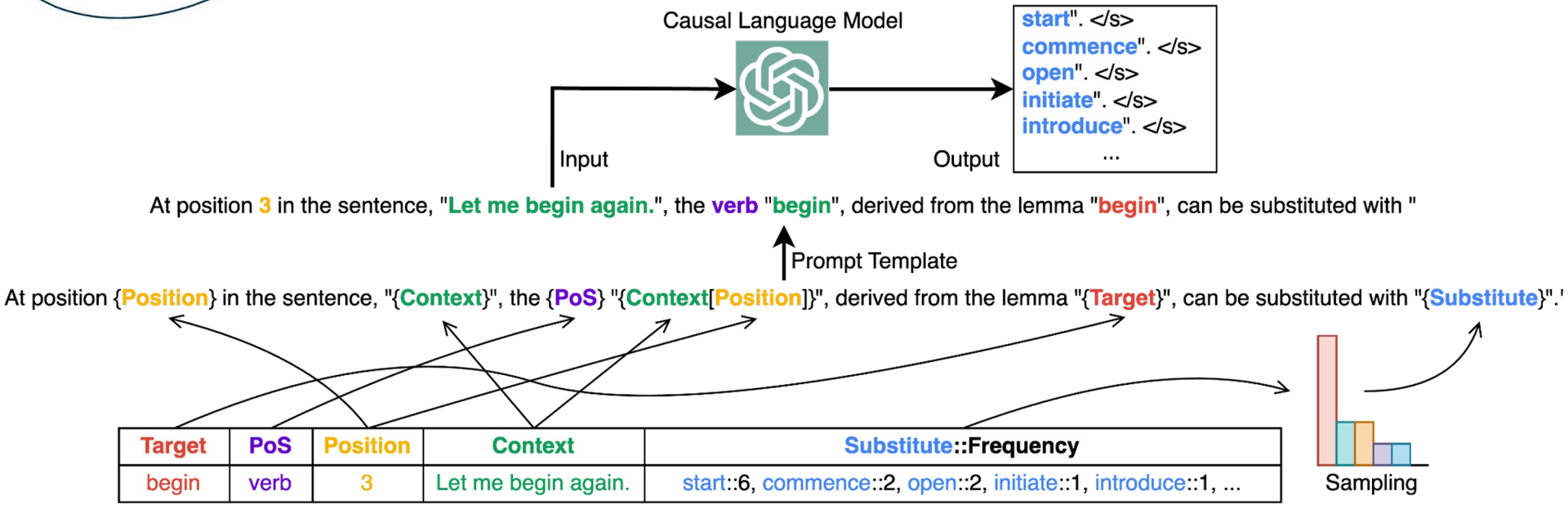
Method – PromptSub
Lexical Substitution via Prompt-aware Fine-tuning

FreqSub exploits the frequency information associated with each gold **substitute**.

Frequency -> Softmax -> Probability Distribution -> Sampling



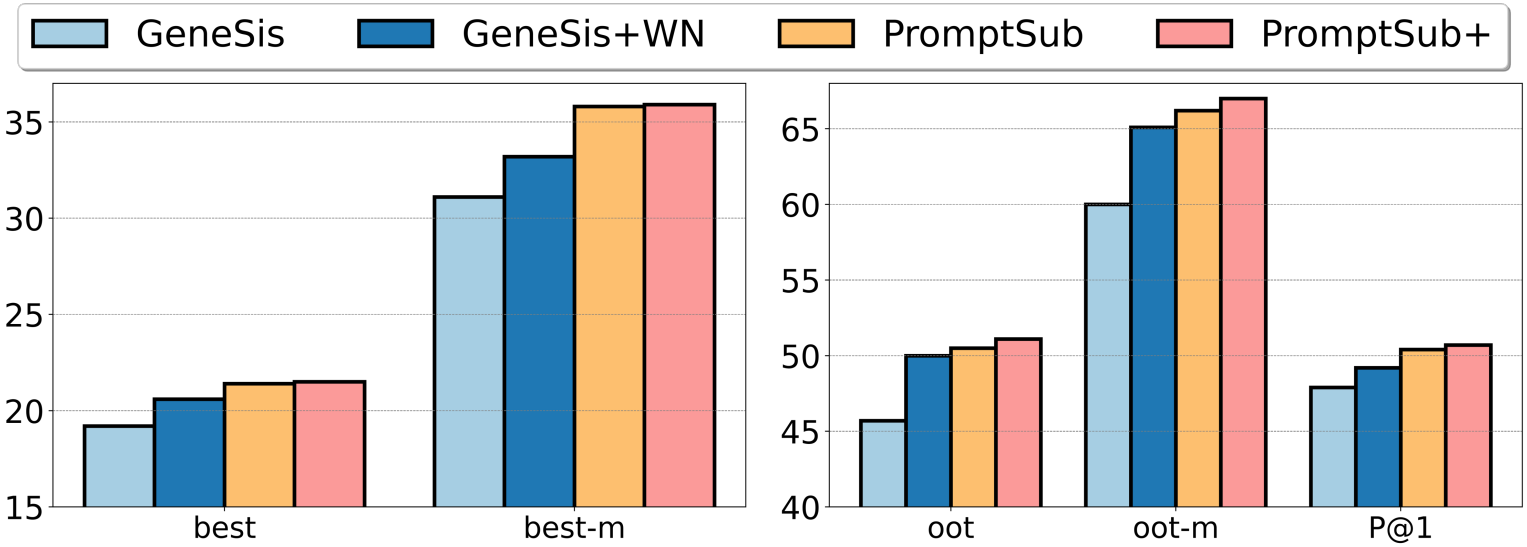
Method – PromptSub
Lexical Substitution via Prompt-aware Fine-tuning



Results – LS07

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

GeneSis+WN relies on external resources from WordNet.



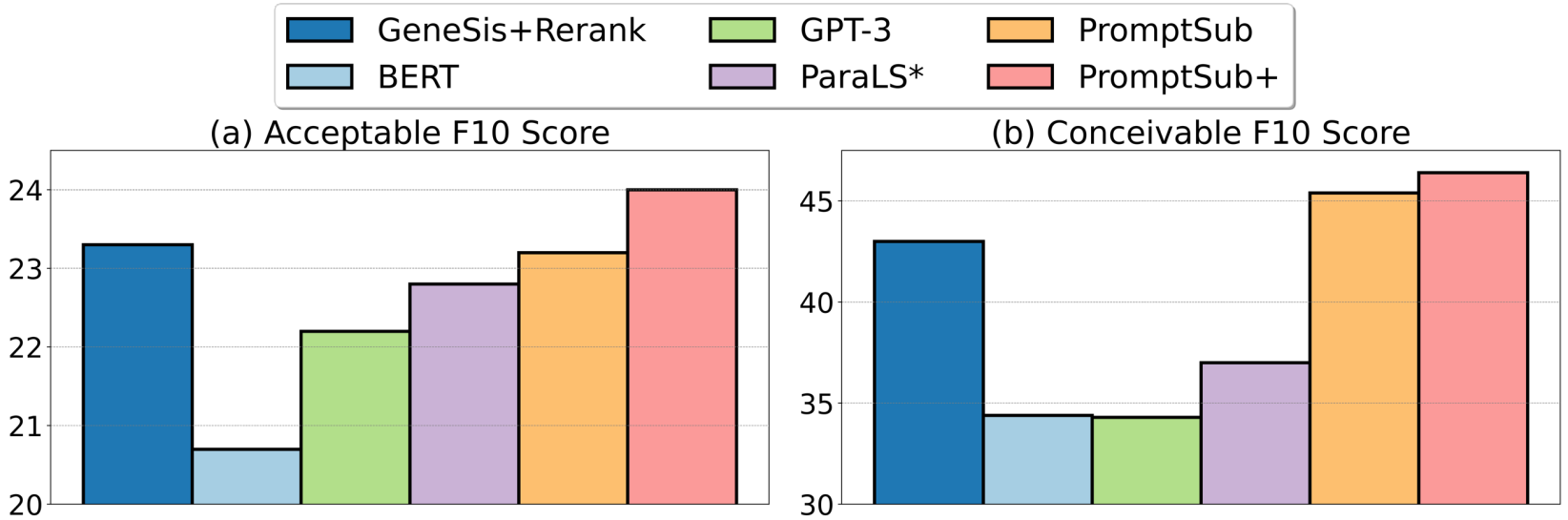
LS07 (Lacerra et al., 2021)

Train: ColnCo and TWSI; Test: Semeval-2007 Task 10

Results – LS21

PromptSub and **PromptSub+** take GPT-2 Medium as its backbone.

GeneSis+Rerank incorporates post-processing to refine its results.

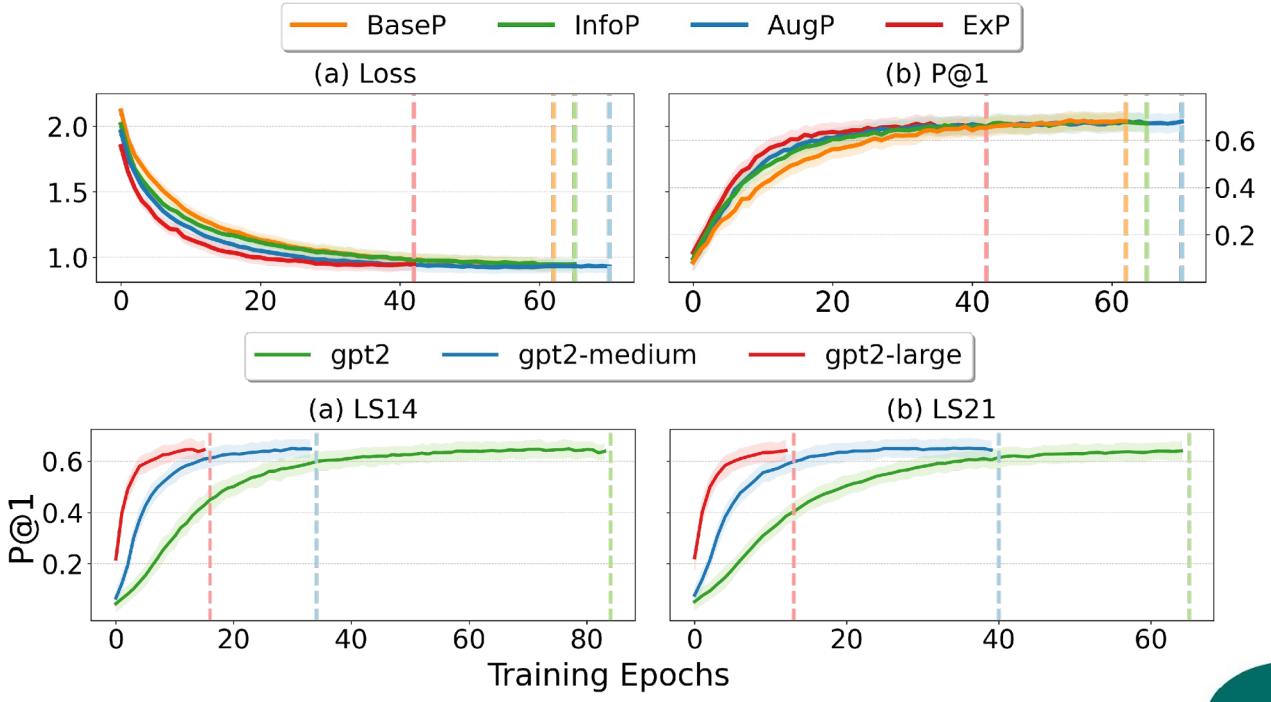


LS21 (Lee et al., 2021)

Train: SWORDS Train, LST and TWSI; Test: SWORDS Test

Scalability

ExPrompt retrieves WordNet synsets for RAG, resulting in **lower** loss, **improved** P@1, and **earlier** convergence.



Conclusion

We have presented **PromptSub**, a framework reducing LST to CLM.

- **Bridges the gap** between pre-training and fine-tuning.
- Takes advantage of **greater** model capacity.
- Leverages a **broad** array of resources.
- Benefits from **external** knowledge through **RAG**.
- Establishes a new overall **state of the art**, particularly LS21.

We expect to extend our approach to other semantic tasks in the future.

github.com/ShiningLab/PromptSub

Reference

Instance:	let me begin again.
BaseP:	the “begin” in the sentence “let me begin again.” can be substituted with “start”.
InfoP:	at position 3 in the sentence, “let me begin again.”, the verb “begin”, derived from the lemma “begin”, can be substituted with “start”.
AugP (Train):	at position 3 in the sentence, “let me begin again.”, the verb “begin”, derived from the lemma “begin”, can be substituted with “start”, “commence”, “open”, “bring about”, “carry on”, “initiate”, “introduce”, “originate”, “restart”, “try”.
AugP (Test):	at position 3 in the sentence, “let me begin again.”, the verb “begin”, derived from the lemma “begin”, can be best substituted with “start”.
ExP (Train):	at position 3 in the sentence, “let me begin again.”, the verb “begin”, derived from the lemma “begin” with synonyms “commence”, “get”, “get down”, “lead off”, “set about”, “set out”, “start”, “start out”, can be substituted with “start”, “commence”, “open”, “bring about”, “carry on”, “initiate”, “introduce”, “originate”, “restart”, “try”.
ExP (Test):	at position 3 in the sentence, “let me begin again.”, the verb “begin”, derived from the lemma “begin” with synonyms “commence”, “get”, “get down”, “lead off”, “set about”, “set out”, “start”, “start out”, can be best substituted with “start”.