Computational Semantics via Lexical Concepts

Ning Shi, **Jai Riley**, Bradley Hauer, Grzegorz Kondrak

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

Outline

This presentation will cover our four recent publications on computational semantics, which address the following research questions:

Q1: Can lexical gaps be detected using machine translation? (ACL 2024)

Q2: How can synonyms be generated for words in context? (*SEM 2024)

Q3: What is the link between semantic similarity and relatedness? (SemEval 2024)

Q4: Can a natural language inference (NLI) model detect paraphrases? (*SEM 2024)

Our Theoretical Assumptions

The 4 papers share a theoretical framework that we developed in our group.

- Lexical **concepts** are discrete units of lexical meaning.
- Each **word sense** corresponds to a distinct concept.
- Each content word in a text expresses **exactly one concept**.
- Words share a **wordnet synset** (synonym set) if and only if they can express the same concept.
- If a word and its translation in some sentence are found in a **bilingual dictionary** then they express the same concept.

Translation-based Lexicalization Generation and Lexical Gap Detection: Application to Kinship Terms

Senyu Li, Bradley Hauer, **Ning Shi**, Grzegorz Kondrak

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

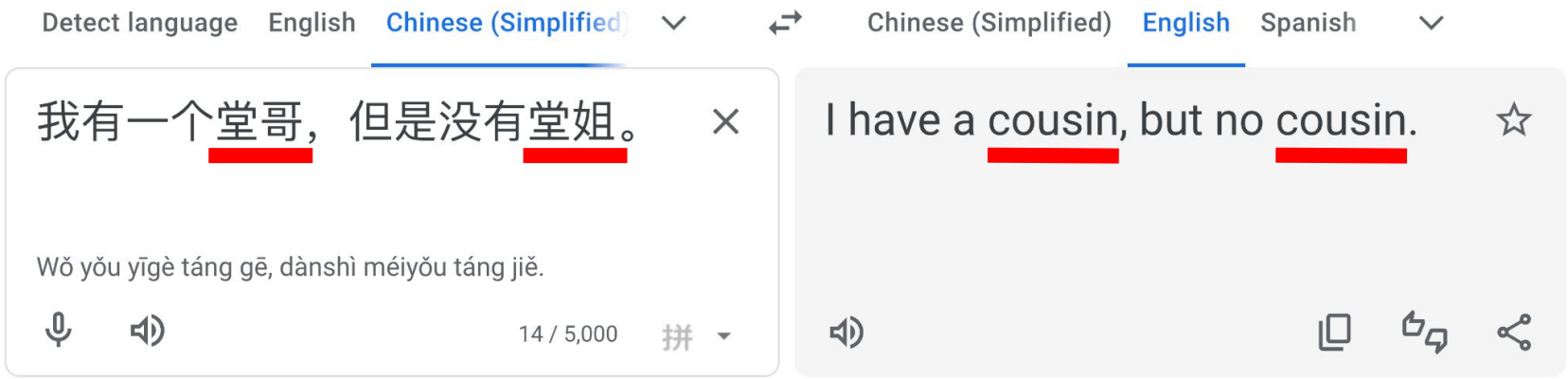
In Proceedings of ACL 2024

An Error Case: Google Translate

堂哥 “elder son of father’s brother” => "**cousin**"

堂姐 “elder daughter of father’s brother” => "**cousin**"

Other powerful translators make similar errors. (DeepL, Baidu, etc.)



Sample Output of ChatGPT

You

Given a word that means [father's younger brother] in Chinese is [叔叔], and a word that means [mother's brother] in Chinese is [舅舅]. Is there a word that means [elder brother] in [English]? If yes, give me that word. If no, say no.

ChatGPT

Yes, the word in English that means "elder brother" is "brother."

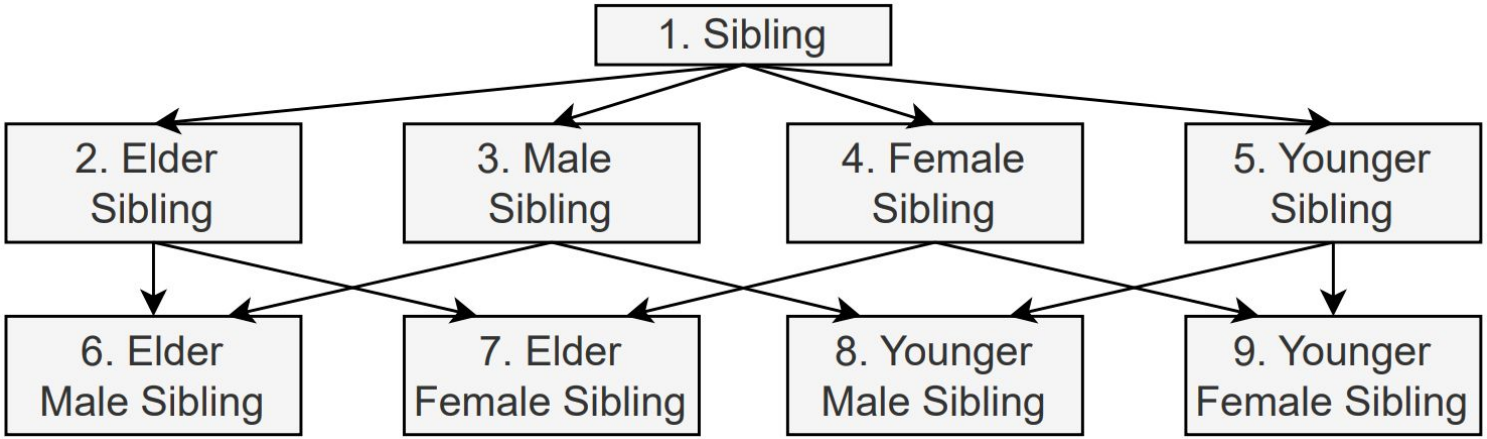
GPT 3.5, Feb. 15, 2024

Concepts

Concept: discrete word meaning

Kinship concepts have clear definitions and hierarchical structure

With well-studied, good gold-standard dataset (Khishigsuren et al, 2022)



Lexicalizations and Lexical Gaps

Lexicalization: a single word which can express (i.e. lexicalize) a concept.

Lexical Gap: a concept that has no lexicalization in a given language.

Concepts	En	Es	Fr	Ja	Fa	Zh	Pl
1	Sibling	ø	fratrie	ø	ø	同胞	ø
2	ø	ø	ø	ø	ø	ø	ø
3	Brother	hermano	frère	ø	برادر	兄弟	brat
4	Sister	hermana	sœur	ø	خواهر	姐妹	siostra
5	ø	ø	ø	ø	ø	ø	ø
6	ø	ø	ø	兄さん	ø	哥哥	ø
7	ø	ø	ø	姉ちゃん	ø	姐姐	ø
8	ø	tato	ø	おとうと	ø	弟弟	ø
9	ø	ø	ø	いもうと	ø	妹妹	ø

Data from *Using Linguistic Typology to Enrich Multilingual Lexicons: the Case of Lexical Gaps in Kinship* (Khishigsuren et al, 2022)

Task definition: LexGen and LexGap

LexGen: Lexicalization Generation

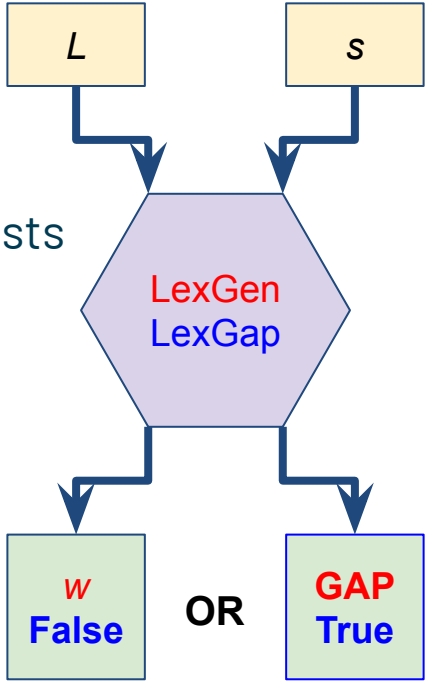
- Input: language *L*, concept *s*
- Output: word *w* in *L* s.t. *w* lexicalizes *s*,

  OR a special token **GAP** indicating that no such *w* exists

LexGap: Lexical Gap Detection

- Input: language *L*, concept *s*
- Output: True if no word in *L* lexicalizes *s*, False otherwise.

LexGen(*L*,*s*) = **GAP** if and only if LexGap(*L*,*s*) = True



Our Work

Problem: How to identify concept lexicalizations and lexical gaps efficiently?

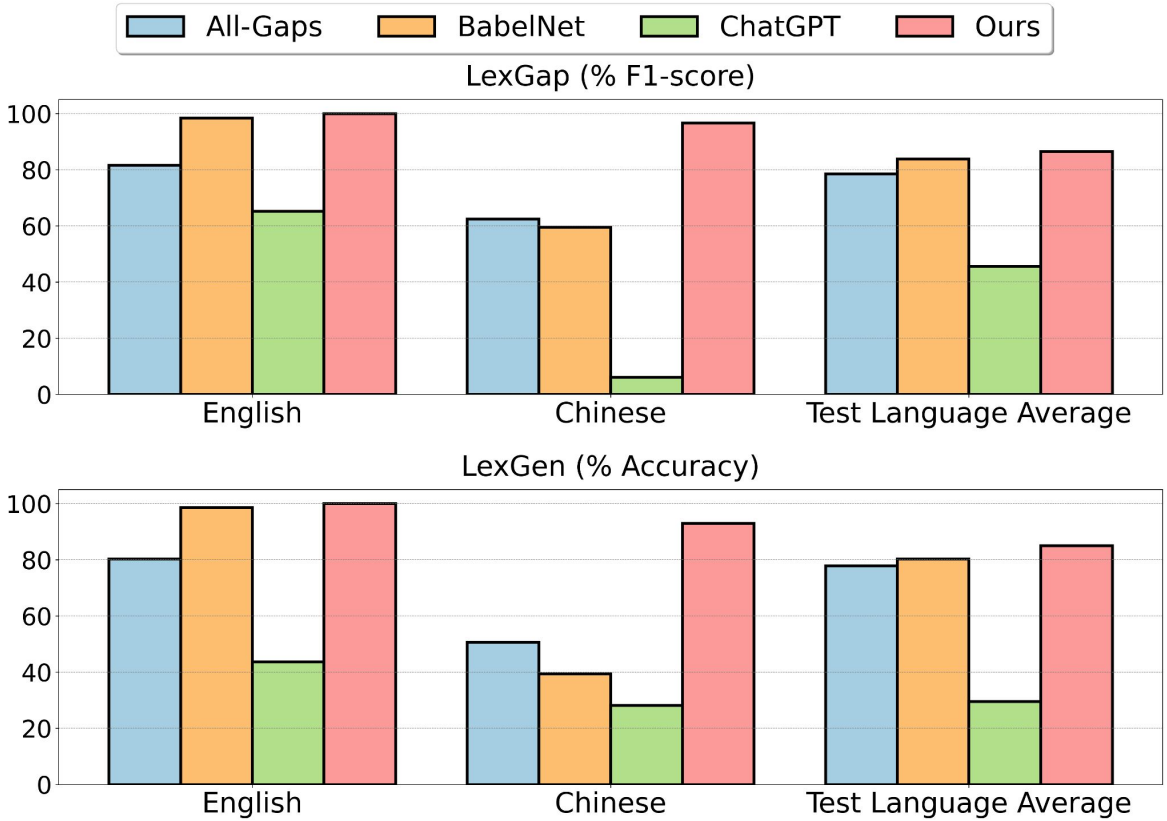
Idea: If a concept is an exclusive disjunction of its hyponym concepts then all three concepts should have different lexicalizations.

Method:

1. Generate a candidate lexicalization for each concept by translating an unambiguous lexicalization into the target language in the context of the concept gloss.
2. Then filter out incorrect translations using the above idea.

Results: Empirical evaluations demonstrate that our approach yields higher accuracy than BabelNet and ChatGPT.

Results


All-Gaps
BabelNet
ChatGPT
Ours
LexGap (% F1-score)
100
80
60
40
20
0
English
Chinese
Test Language Average
LexGen (% Accuracy)
100
80
60
40
20
0
English
Chinese
Test Language Average

Conclusion

- Novel translate-and-filter method for:
  - Generating **lexicalizations**
  - Detecting **lexical gaps**
- Grounded in linguistic theory, with clear definitions and and propositions
- Leverages translation and hypernym/hyponym relations
- Future work: Beyond kinship to other domains

github.com/UAlberta-NLP/KinshipAutoLex

Lexical Substitution as Causal Language Modeling

Ning Shi, Bradley Hauer, Grzegorz Kondrak

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

In Proceedings of *SEM 2024

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." → MLM → "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: 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, ...

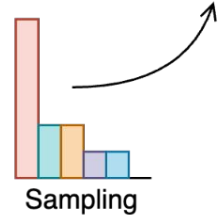
Method – PromptSub
Lexical Substitution via Prompt-aware Fine-tuning

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

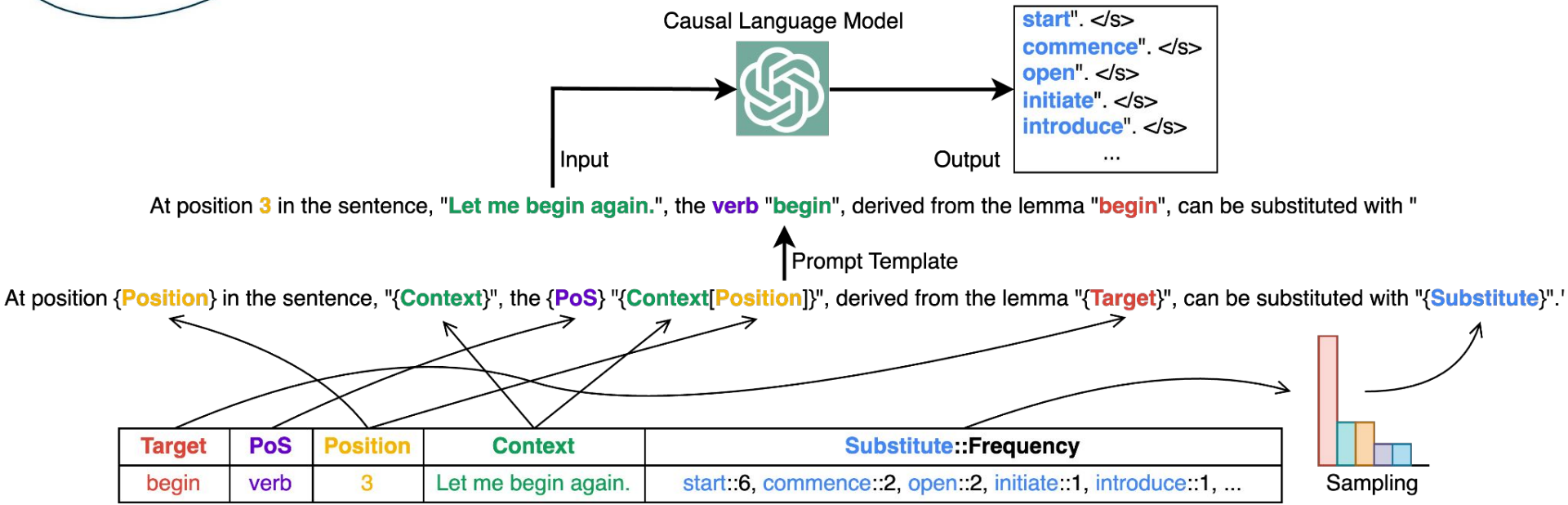
Frequency -> Softmax -> Probability Distribution -> Sampling

At position {Position} in the sentence, "{Context}", the {PoS} "{Context[Position]}", derived from the lemma "{Target}", can be substituted with "{Substitute}".'

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



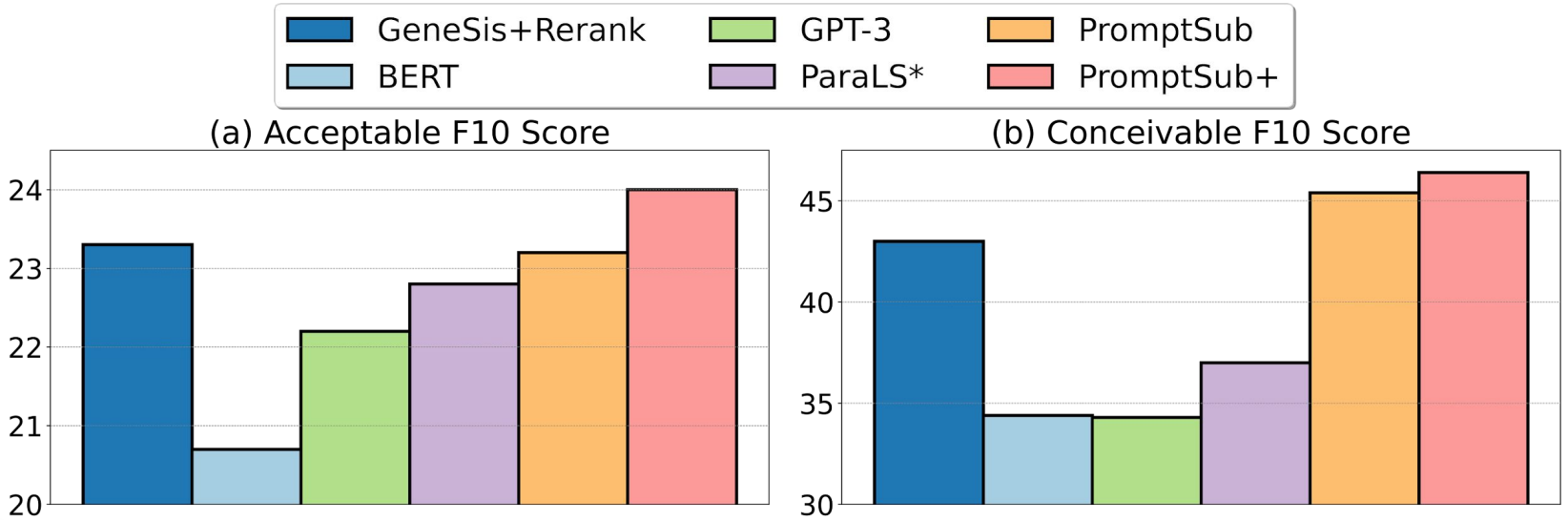
Method – PromptSub
Lexical Substitution via Prompt-aware Fine-tuning



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

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.
- 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

UAlberta at SemEval-2024 Task 1: A Potpourri of Methods for Quantifying Multilingual Semantic Textual Relatedness and Similarity

Ning Shi, Senyu Li, Guoqing Luo, Amirreza Mirzaei, Ali Rafiei
Jai Riley, Hadi Sheikhi, Mahvash Siavashpour
Mohammad Tavakoli, Bradley Hauer, Grzegorz Kondrak

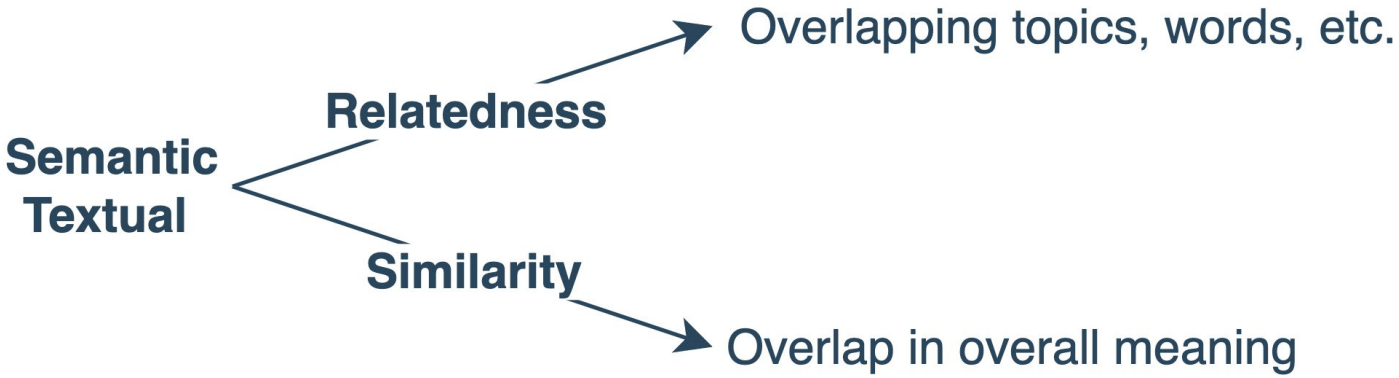
Alberta Machine Intelligence Institute (Amii)
Department of Computing Science
University of Alberta, Edmonton, Canada

In Proceedings of SemEval 2024
(🏆 1st Place on Track A English among all submitted systems)

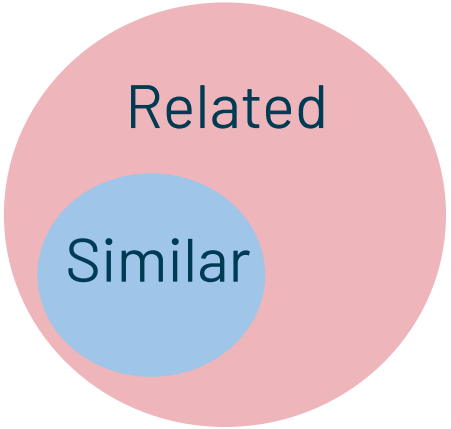
Semantic Textual Relatedness (STR) and Similarity (STS)

STR measures the **degree of commonality** between pairs of sentences.

STS measures the degree in which pairs of sentences are **close in meaning.**



Hypothesis 1



Similarity is a special case of relatedness.

For example:

And in the United States, **we're considered** Mexican.

And in the United States, **we are considering** Mexicans.

High relatedness but **low** similarity.

*SemEval 2017 Task 1 Track 4b

Hypothesis 2

Relatedness and similarity are **preserved under translation.**

It is better **known as a walk** . **It is** also **known as a walk** .	0.88	
It is better **known as a walk** . También **se le conoce como paseo** .	0.88	Spanish
It is better **known as a walk** . **Dit staan** ook **bekend as 'n stap** .	0.88	Afrikaans

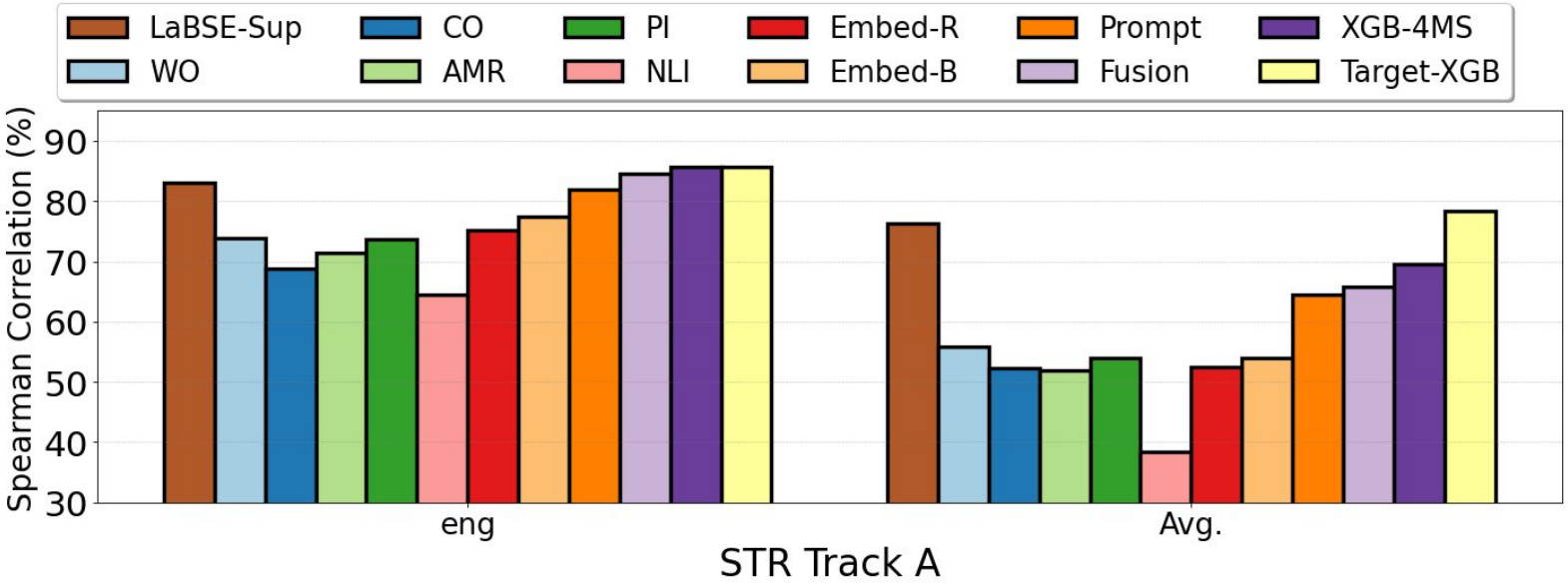
Methods

Explicit Semantic	Extrinsic	Distributional	Large Language Models
• Word Overlap **(WO)** • Concept Overlap **(CO)** • Abstract Meaning Representation **(AMR)**	• Paraphrase Identification **(PI)** • Natural Language Inference **(NLI)**	• **Embed-B:** BERT • **Embed-R:** RoBERTa	• **Prompt:** GPT-3.5 • **Fusion:** SBERT • **Fine-tune:** T5, GPT2, RoBERTa, MPNet

Our **best results** are reported from a **XGBoost** (Chen & Guestrin, 2016) regression ensemble system involving the **4 fine-tuned models.**

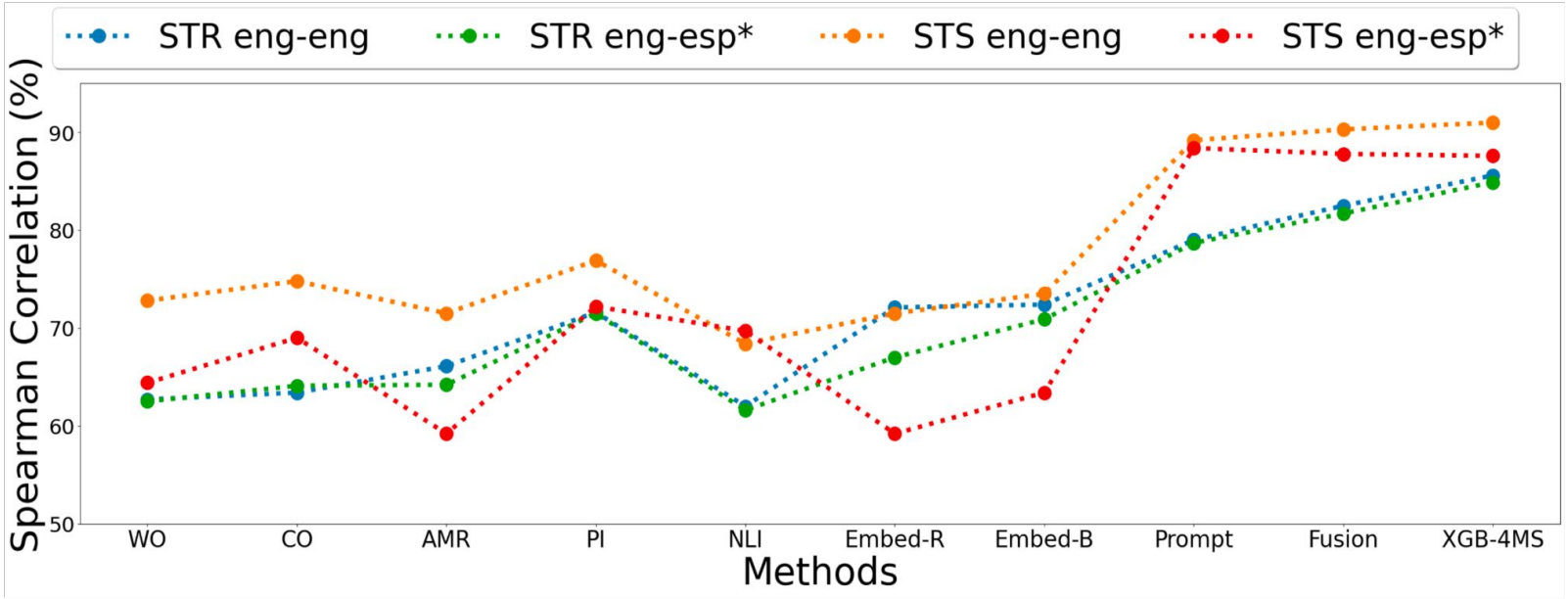
STR Results

Achieved SOTA results for English.



STR vs STS Results

High correlation between performance of methods on STR and STS datasets.



Conclusions

- **Ensembled a variety of methods** on two sentence-level semantic tasks in mono-lingual and cross-lingual conditions.
- Achieved **SOTA results for English** and **top 3** performance for **16** of the language/track settings.
- Provided evidence for **two hypotheses**:
  1. Semantic **similarity is a special case of semantic relatedness.**
  2. **Both** similarity and relatedness are **preserved under translation.**

github.com/UAlberta-NLP/SemEval2024-STR

Paraphrase Identification via Textual Inference

Ning Shi, Bradley Hauer, **Jai Riley**, Greg Kondrak

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

In Proceedings of *SEM 2024

Natural Language Inference

Natural **L**anguage **I**nference (NLI) involves three labels that describe the relationship between two sentences.

Entailment, Contradiction, Neutral

For example:

S_1: "This man is surfing."
S_2: "A man is on water."

Surfing: an aquatic activity or website browsing?

Paraphrase Identification

Paraphrase **I**dentification (PI) is the task of deciding whether two sentences convey the same meaning.

Hypothesis:

Paraphrasing corresponds to **bidirectional textual entailment**.

Prior work:

- A blend of modules complicates the analysis
- Bias to traditional PI methods
- Lacks any theoretical formalization

Our Contributions

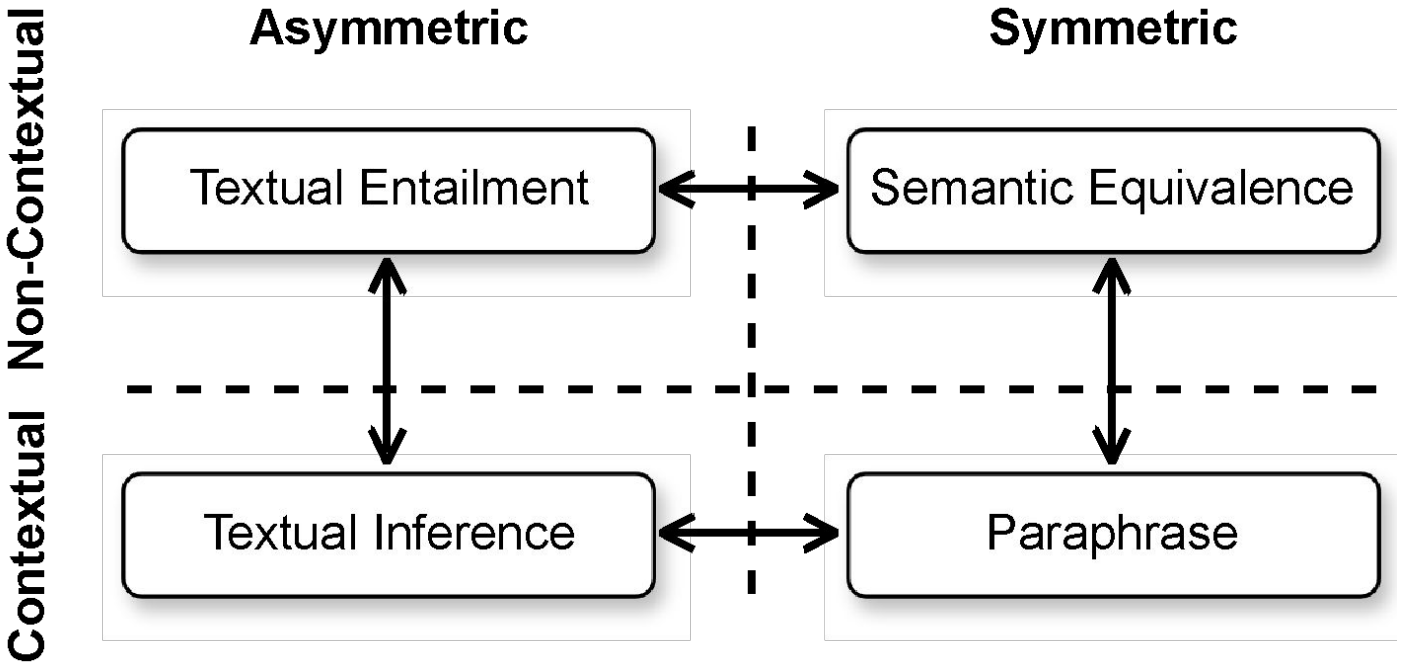
We present the **first theoretical formalization** implying a **practical reduction of PI to NLI** (PI2NLI), validated by fine-tuning an NLI model for PI.

- A theoretical task reduction showing how PI can be reduced to NLI.
- Extensive evaluation across zero-shot and fine-tuning.
- We found fine-tuned NLI models can **outperform** dedicated PI models on PI datasets.

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

(Hauer and Kondrak, 2023)

Sentence-level Relations



Equivalence and Paraphrasing

Semantic **Eq**uivalence relation**, SEQ(S_1, S_2)** :=
"the sentences S_1 and S_2 convey the same meaning"

PR(C, S_1, S_2) := "the sentences S_1 and S_2 convey the same meaning given the context C"

The relationship in between:

$$\mathbf{SEQ(S_1, S_2) \Leftrightarrow \forall C : PR(C, S_1, S_2)}$$

Example:

S_1: "We must work hard to win this election."
S_2: "The Democrats must work hard to win this election."

Entailment and Inference

Textual **E**ntailment, $TE(S_1, S_2)$:=
"the sentence S_2 can be inferred from the sentence S_1"

Textual **I**nference, $TI(C, S_1, S_2)$:= "the sentence S_2 can be inferred from the sentence S_1 given the context C"

The relationship in between:

$$\mathbf{TE(S_1, S_2) \Leftrightarrow \forall C : TI(C, S_1, S_2)}$$

Example:
S_1: "This man is surfing."
S_2: "A man is on water."

Proposition

Given context C, sentences S_1 and S_2 are paraphrases if and only if they can be mutually inferred from each other.

Formally:

$$\mathbf{PR(C, S_1, S_2) \Leftrightarrow TI(C, S_1, S_2) \wedge TI(C, S_2, S_1)}$$

Context:

- Context includes common sense and world knowledge.
- In practice, context is embedded in the data distribution.

Data Adaptation

Positive PI instances:

We convert each positive PI instance into two distinct NLI positive instances, one in each direction.

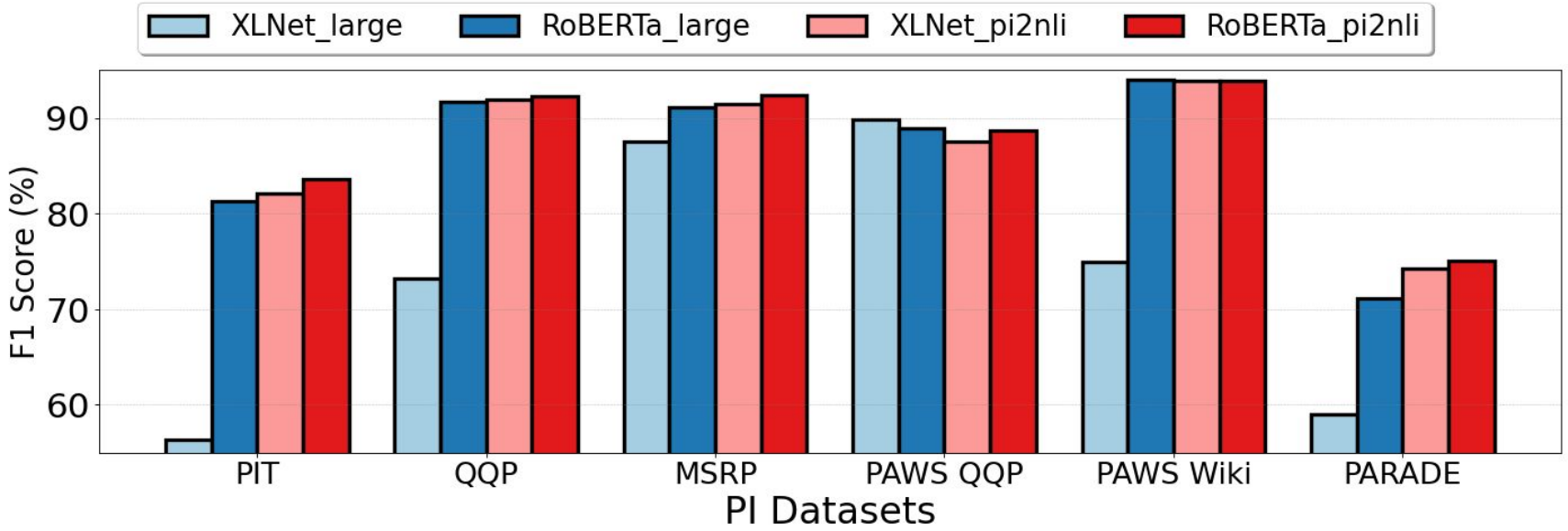
PI(S_1, S_2, True) -> NLI(S_1, S_2, Entailment) **AND** NLI(S_2, S_1, Entailment)

Negative PI instances:

We generate a negative NLI instance (randomly selected as either Contradiction or Neutral) in one randomly selected direction.

PI(S_1, S_2, False) -> NLI(S_1, S_2, Contradiction) **OR** NLI(S_2, S_1, Contradiction)
NLI(S_1, S_2, Neutral) **OR** NLI(S_2, S_1, Neutral)

Results



XLNet_large (Yang et al., 2019); RoBERTa_large (Liu et al., 2019)

PAWS QQP and PAWS Wiki include **adversarial** example created by word scrambling and back-translation.

Conclusion

We have presented PI2NLI, the first attempt to reduce PI to NLI.

- PI can be reduced to NLI **theoretically** and **empirically**.
- Fine-tuned NLI models can **outperform** PI models on PI datasets.
- Applying PI2NLI in a zero-shot setting show the **limitations** in the current PI datasets.

Future:

- Using NLI to refine PI data?
- Using PI models to solve NLI?
- Using one single model to solve both PI and NLI?

github.com/ShiningLab/PI2NLI

Summary of Findings

A1: Machine translation, combined with hypernymy relations, can effectively detects lexical gaps in languages, surpassing BabelNet and ChatGPT (ACL 2024).

A2: Causal language models can generate accurate synonyms in context, outperforming previous SOTA methods (*SEM 2024).

A3: Semantic similarity is a subset of relatedness, and our methods capture both effectively (SemEval 2024).

A4: Natural language inference models can detect paraphrases via a task reduction (*SEM 2024).

Future Work

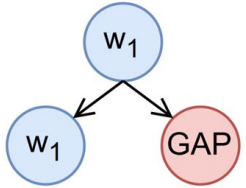
We aim to develop a unified framework for NLP tasks based on semantics, offering linguistic grounding, empirical validation, and resources.

- Construct a taxonomy that connects tasks like contradiction detection (NLI) to word prediction (LM).
- Leverage Translation Identification in pre- and post-processing for Machine Translation systems
- Develop easy-to-implement language-agnostic pipelines for creating multilingual NLP task datasets.

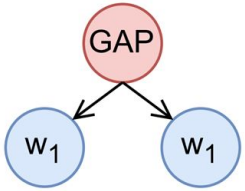
Thanks!
Q & A

Kinship - Theoretical Basis

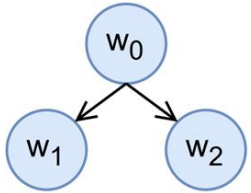
- **Proposition 1:** If a concept P is an exclusive disjunction of its hyponym concepts C1 and C2, expressing P and C1 with the same word w can result in a colloquial contradiction.



- **Proposition 2:** If a concept P is an exclusive disjunction of its hyponym concepts C1 and C2, expressing C1 and C2 with the same word w can result in a colloquial contradiction.



- **Corollary:** If a concept P is an exclusive disjunction of its hyponyms C1 and C2 then all their lexicalizations should be different.



Kinship - Our Method

Generate a candidate lexicalization w for each concept by **translating a seed word.**

Filter using our four-step procedure:

1. **Multi-word filter**: If w is not a single word (e.g. "male cousin"), **return GAP**
2. **Horizontal filter (Proposition 2)**: If w was also generated for a sibling node of s, **return GAP**
3. **Back-translation filter**: If back-translating w does not recover the seed word, **return GAP**
4. **Vertical filter (Proposition 1)**: If w was also generated for a parent node of s, and another child of that parent node has already been tagged as a **GAP**, then **return GAP**

If *w* makes it past the filters, return *w* for LexGen, **False** for LexGap

Kinship - Experimental Setup

Data: Database of Lexical Diversity in Kinship by Khishigsuren et al. (2022)

Translator: Google Translate

Metrics: Accuracy for LexGen, F1 score for LexGap

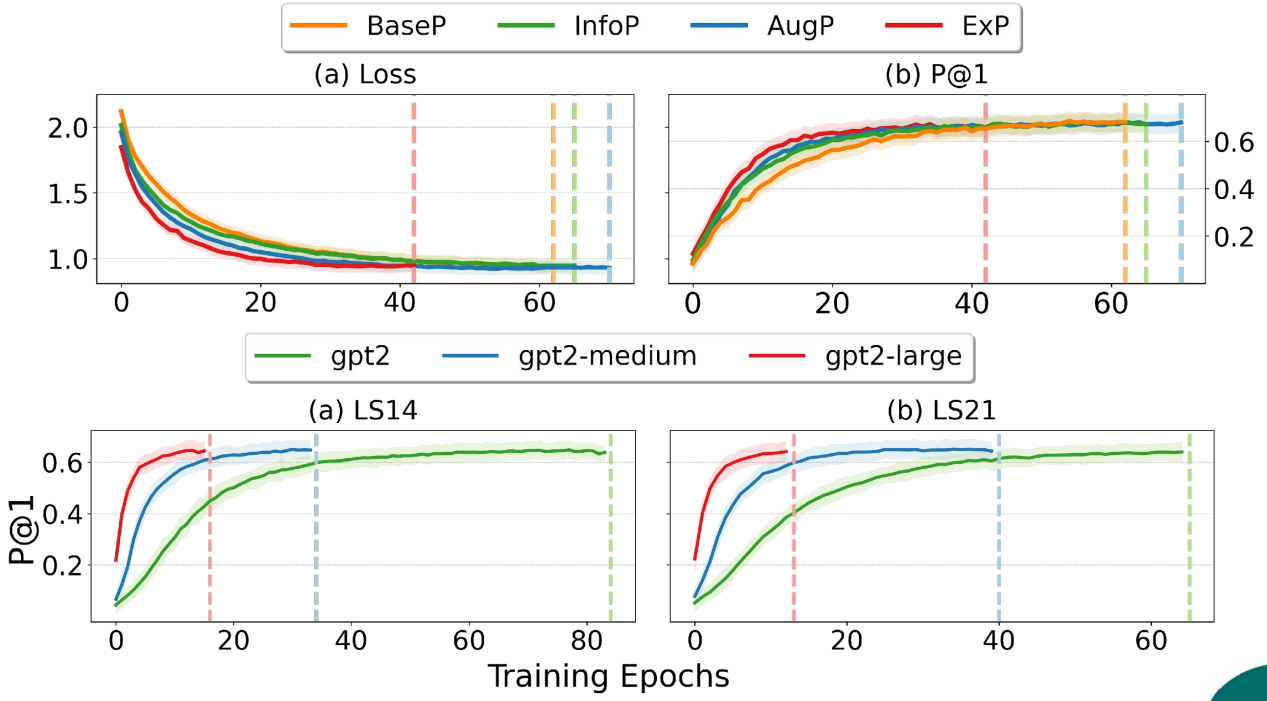
Comparison: All-Gaps, BabelNet 5.1, and ChatGPT w/ GPT-3.5 Turbo

Languages

- Development languages: English, Mandarin, and Persian.
- Test languages: Spanish, Russian, French, German, Polish, Arabic, Italian, Mongolian, Hungarian, and Hindi.

PromptSub - Scalability

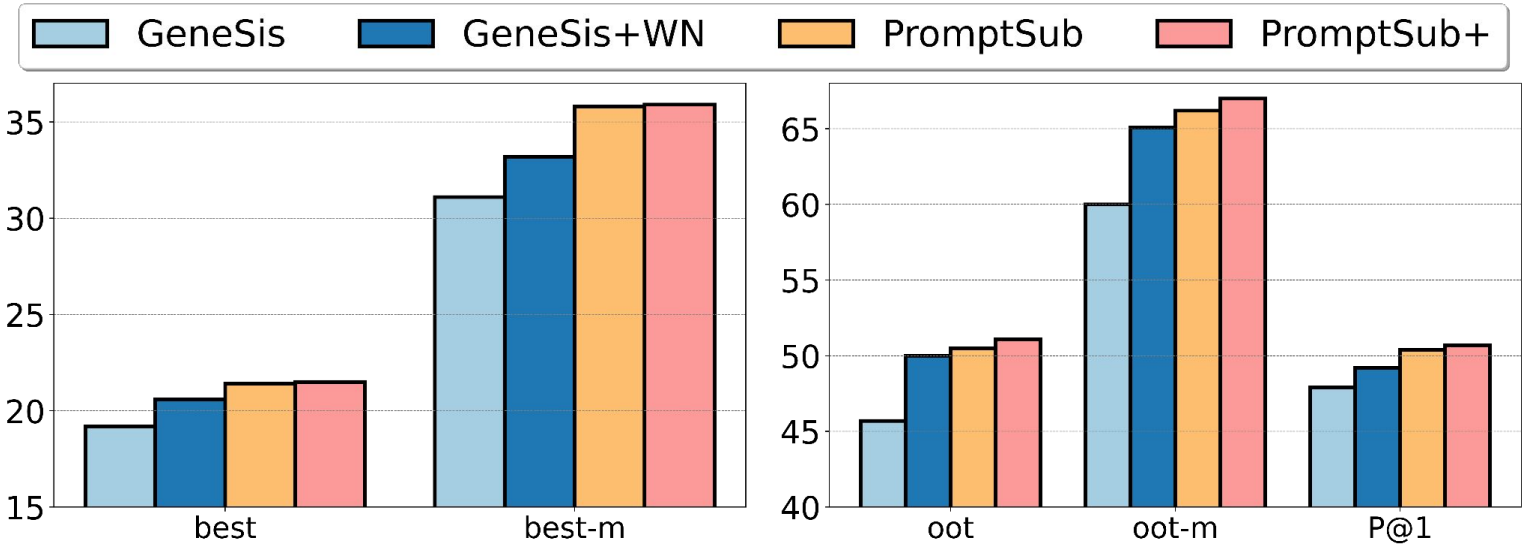
ExPrompt retrieves WordNet synsets for RAG, resulting in **lower** loss, **improved** P@1, and **earlier** convergence.



PromptSub – 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: CoInCo and TWSI; Test: Semeval-2007 Task 10

PromptSub – Examples

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".

SemEval - Methods

Explicit Semantic	Extrinsic	Distributional	Large Language Models
Create and compare semantic representations of each inputted sentence	Use the output of systems designed for other semantic tasks	Create and compare embeddings from PLMs	Prompting or combining multiple model outputs

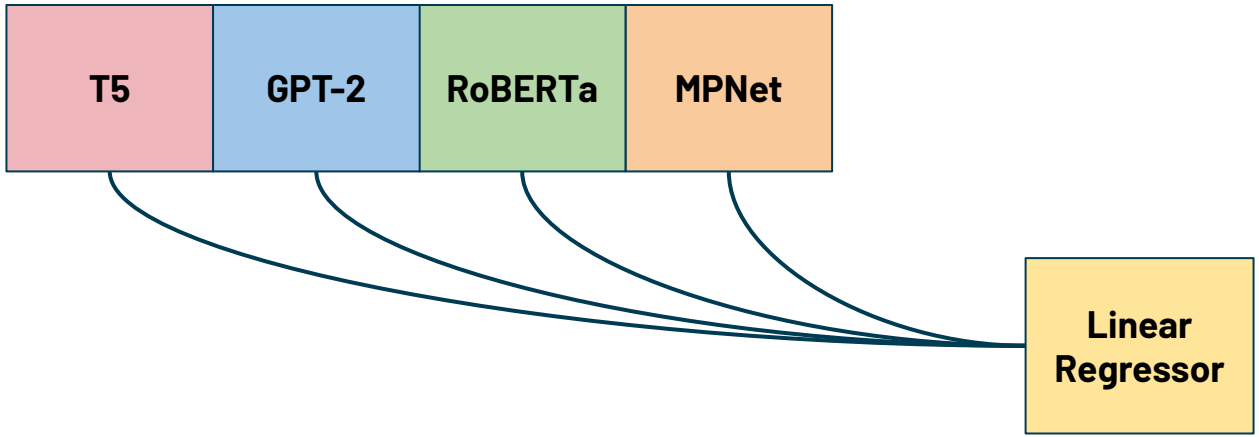
SemEval - Methods

Explicit Semantic	Extrinsic	Distributional	Large Language Models
WO: Python Libraries **CO:** AMuSE-WSD **AMR:** Sapienza API	**PI:** RoBERTa & fine-tuned classifier **NLI:** RoBERTa with NLI Classifier	**Embed-B:** BERT **Embed-R:** RoBERTA	**Prompt:** ChatGPT **Fusion:** Open-source LLMs **Fine-tune:** T5, GPT2, RoBERTa, MPNet

SemEval - Ensemble

Our **best results** are reported from a **regression ensemble system** involving the **4 fine-tuned models.**

- Treat **each score** as a **feature in a linear regressor.**



SemEval - STR and STS

When I tried again, I was able to juggle.
When I went back to it, I was able to juggle!

High relatedness
High similarity

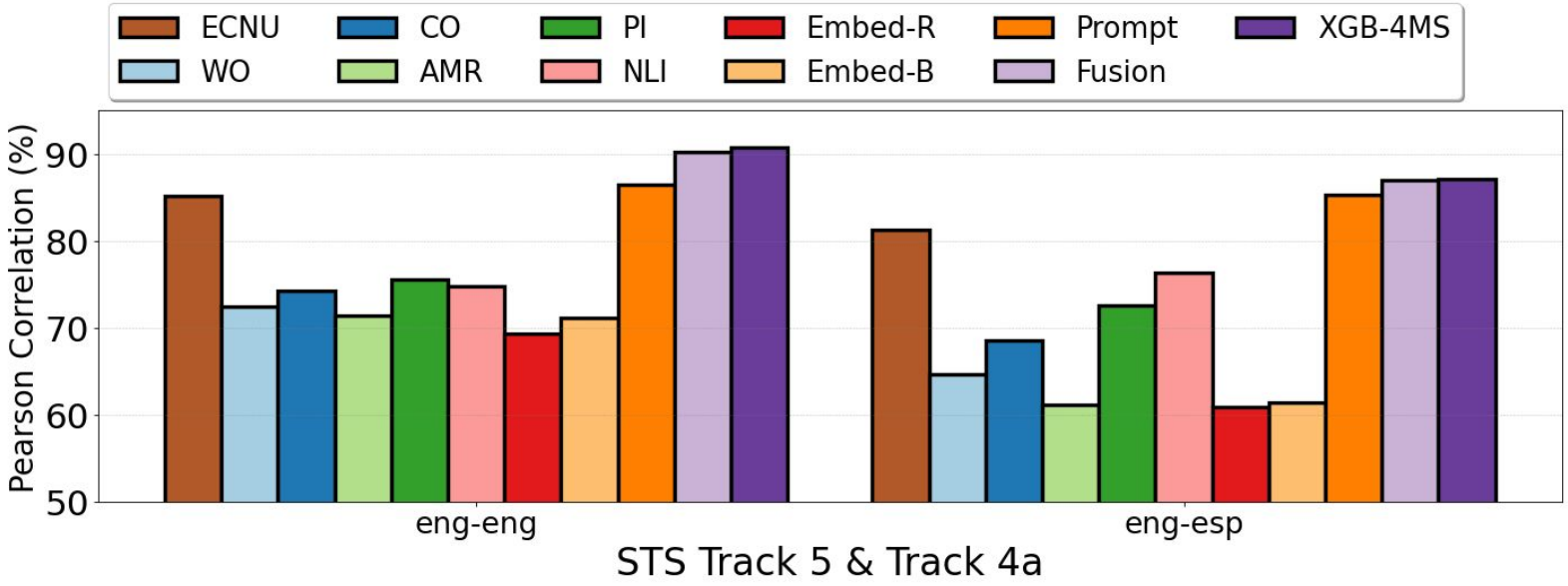
Old car driving down the road.
Two old women enjoying at a gathering.

Low relatedness
Low similarity

*SemEval 2024 Task 1 English Train

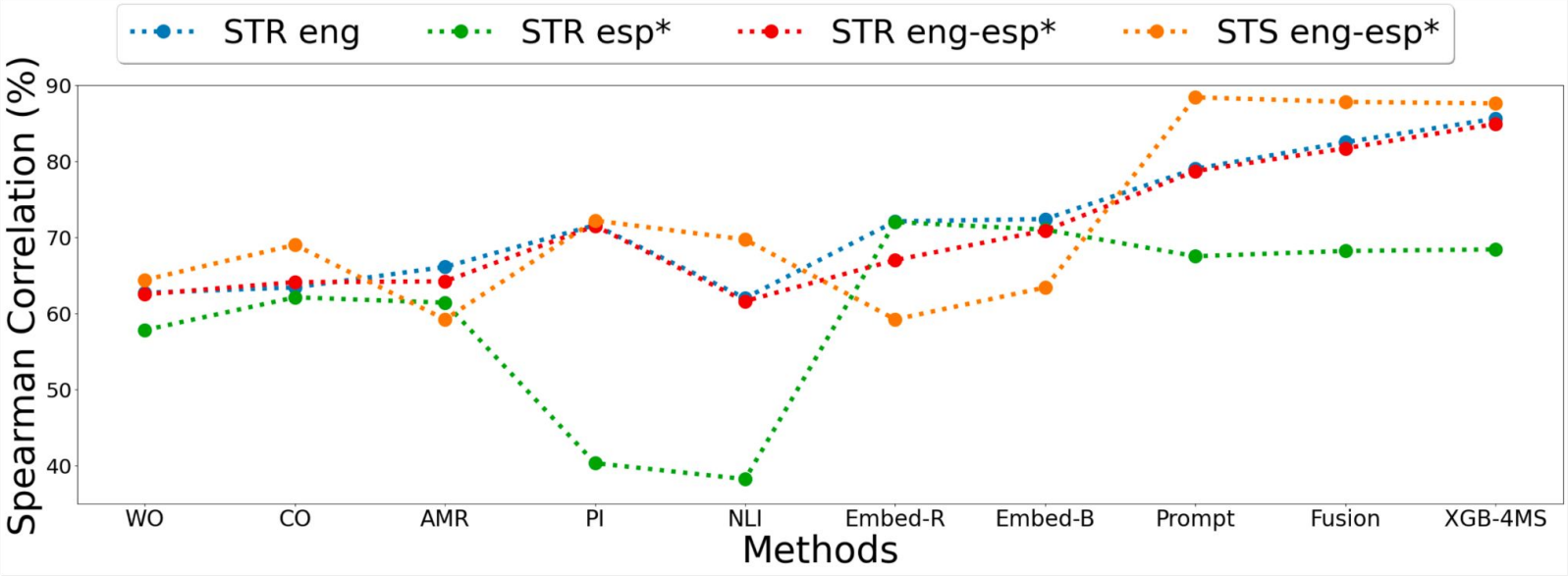
SemEval - STS Results

STS dataset from SemEval 2017 Task 1 with ECNU being the best recorded method.

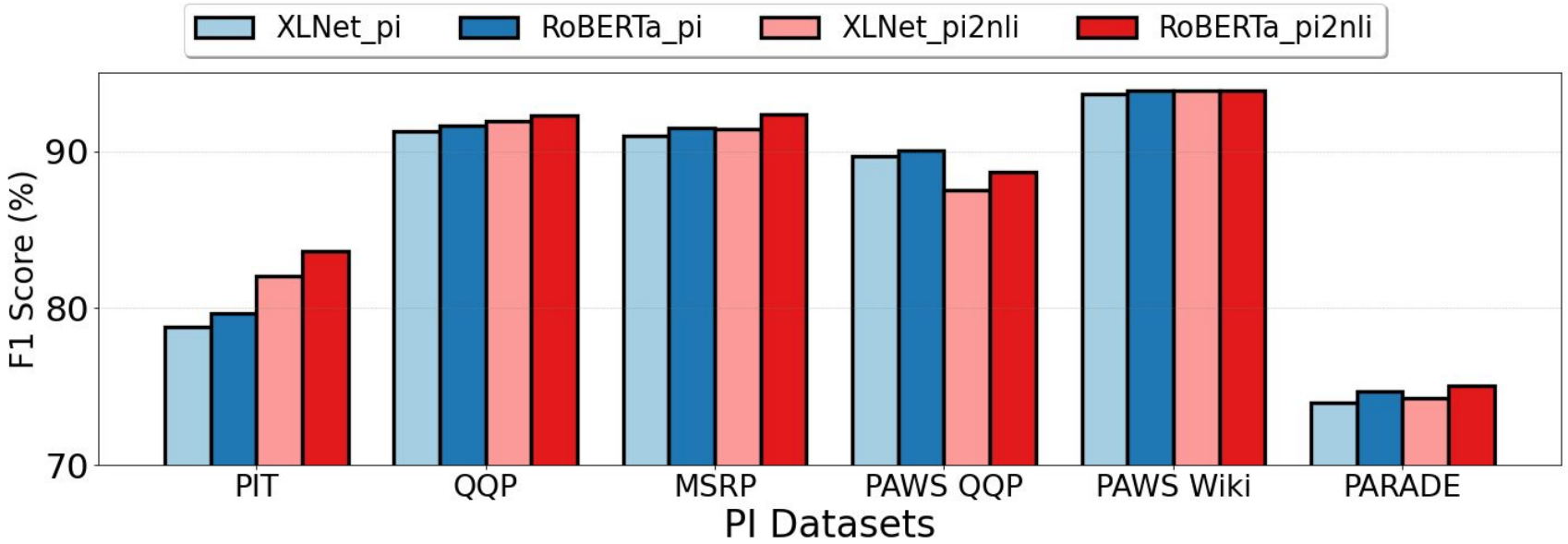


SemEval - Mono-Lingual vs Cross-Lingual

Ning: Can we also have a conclusion sentence here?



PI2NLI - Ablation



XLNet_pi, RoBERTa_pi (Nie et al., 2020)

The **same** language models with classification heads initialized from **scratch**.

PI2NLI - Analysis

A paraphrase identified in one dataset might **NOT** necessarily be considered a valid paraphrase in the other.

We view this **adjustment** as the process of how models learn the **context** inherent in each PI dataset.