Unsupervised Syntactic Parsing via Max. Semantic Information

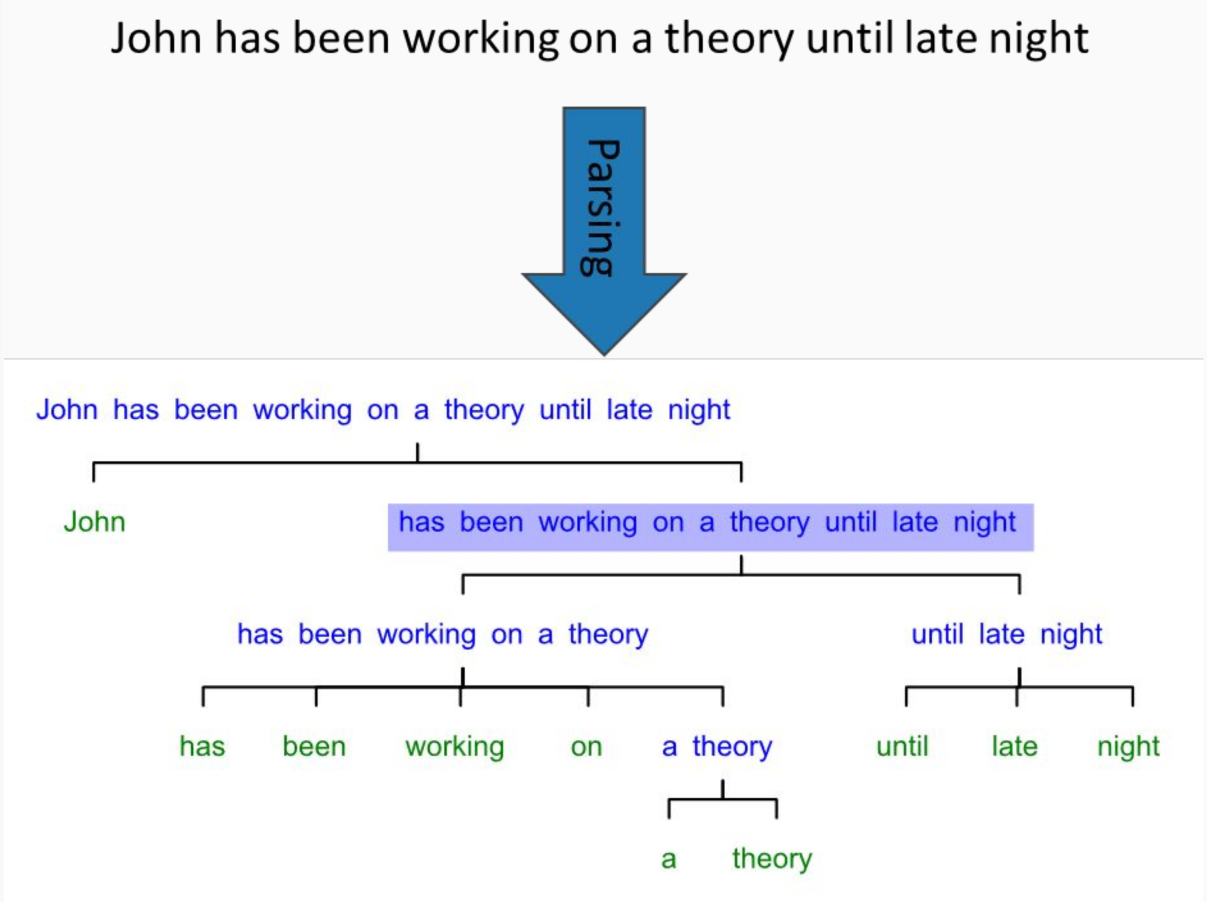
Junjie Chen

Unsupervised Syntactic Parsing

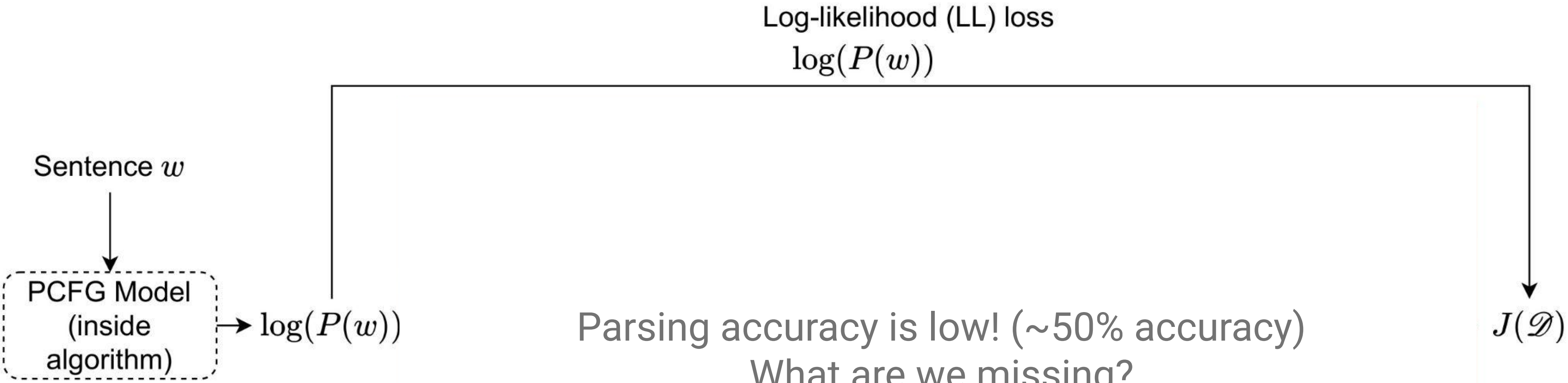
Parsing: Finding a tree structure where sub-units (e.g., substring/constituents) carry significant **semantic information**.

Why Unsupervised Parsing:

- An **unsupervised inference** of the semantic structure.
- A data-driven study of information structure in natural language



Beyond Language Modeling?

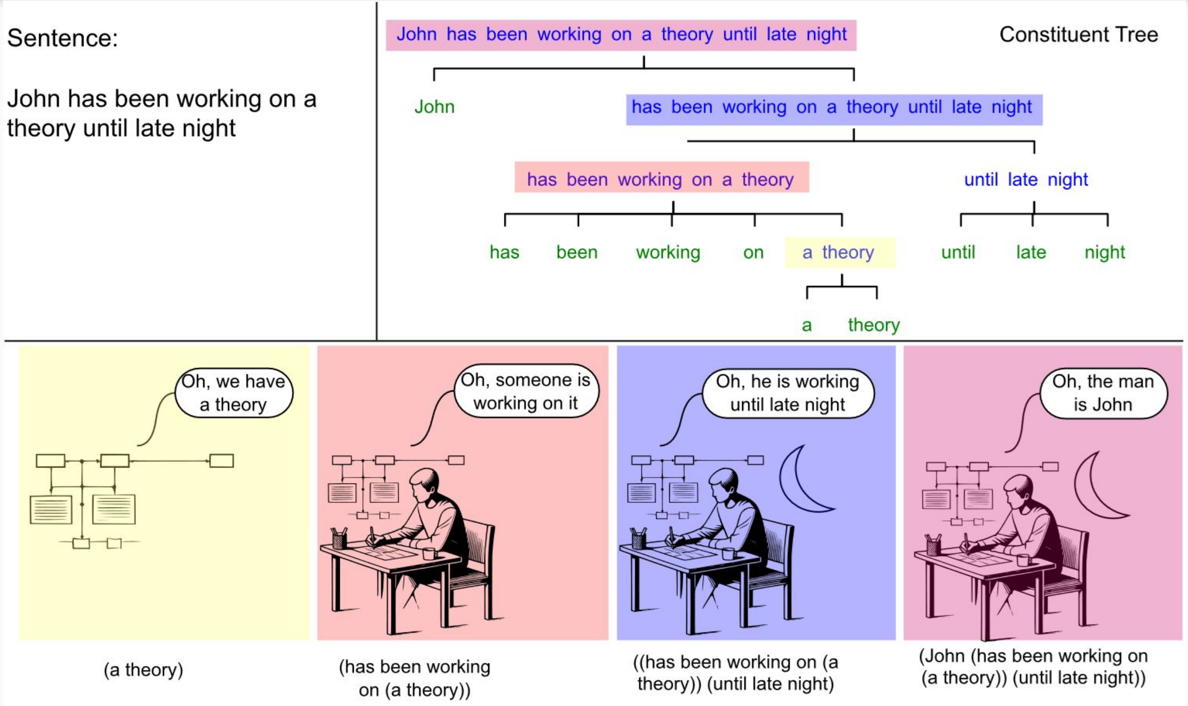


Parsing accuracy is low! (~50% accuracy)
What are we missing?

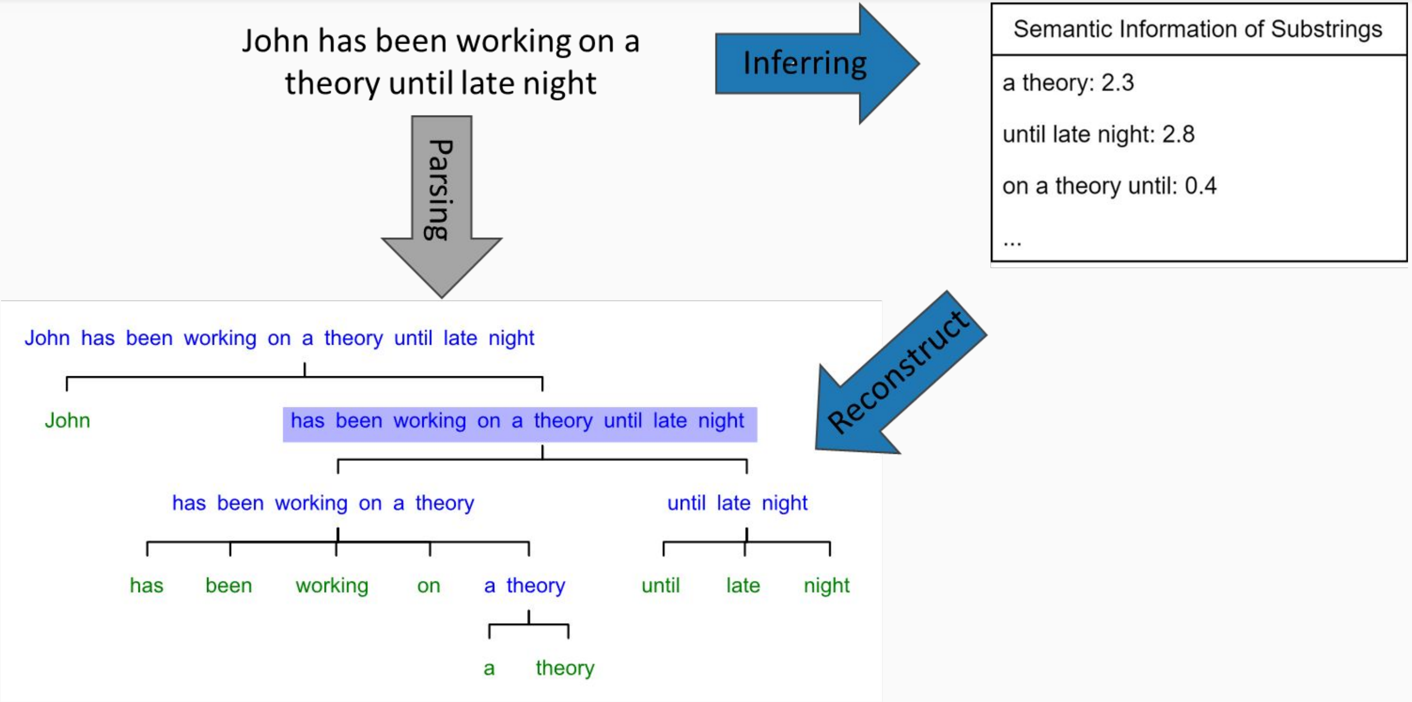
Failing to Model Communication Message (Semantic Information)

Syntactic structure encodes semantic information as messages in communication

Parsing by maximizing semantic information



Parsing by Max. Semantic Information



Parsing by Max. Semantic Information

Estimating Substring Semantic Information $I(t, Sem(x))$ via **paraphrasing**

- Paraphrasing exposes semantic-driven word collocations
- Estimating substring semantic information using bag-of-substrings model and TF-IDF

Reinforcement learning to encourage high SemInfo predictions

- Training a parser by maximizing $I(t, Sem(x))$
- Enforcing Tree-constraint through RL

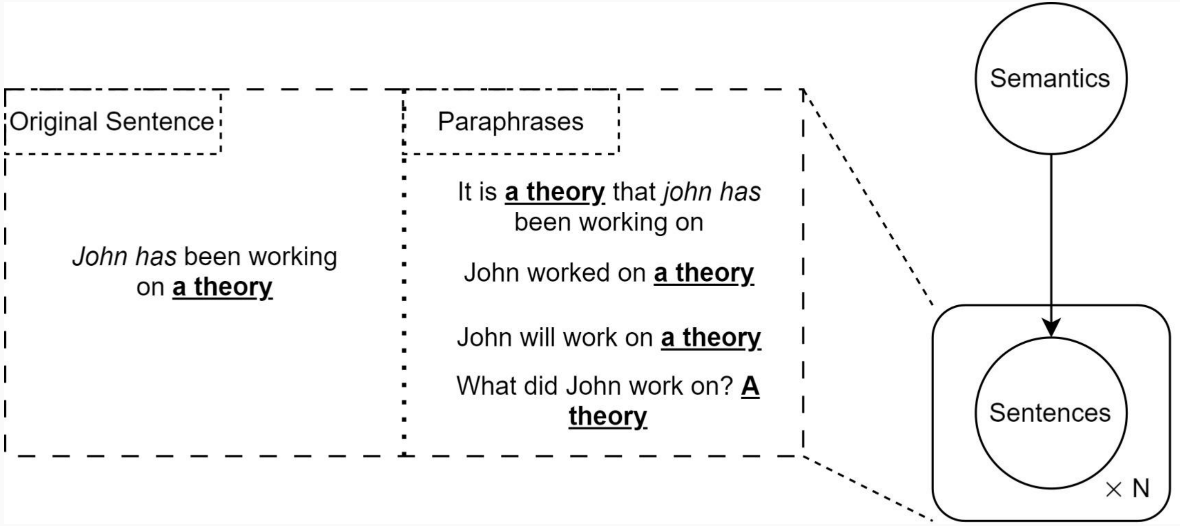
Frequency in Paraphrases Reveals Semantic-driven Word Collocations

Paraphrasing:

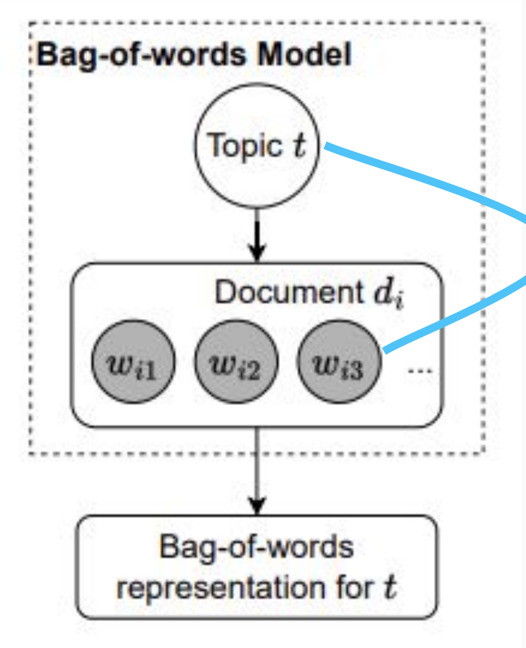
- Preserving semantic-driven collocations (**freq**. ↑)
- Breaking superficial collocation (**freq**. ↓)

Word collocations:

- A substring (i.e., a node in syntactic tree)

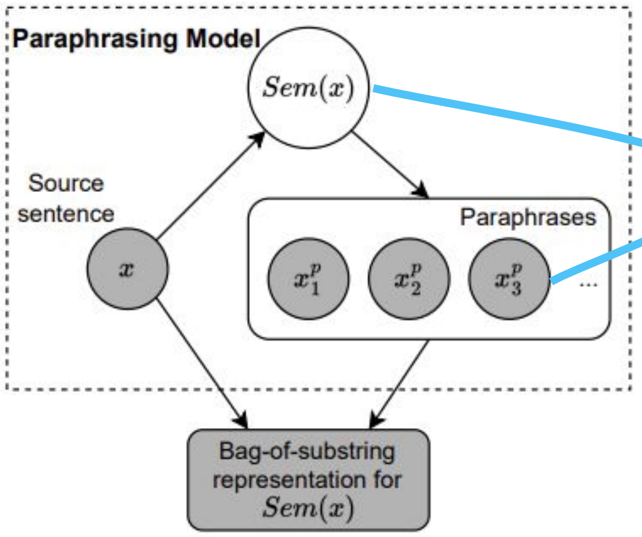


Estimating Substring Semantic Information using TF-IDF



$$I(w_{i,j}, d_i) = \underbrace{F(w_{i,j}, d_i)}_{\text{Word Frequency in document}} \underbrace{\log \frac{|\mathcal{D}|}{|d' : w_{i,j} \text{ is a word in } d'|}}_{\text{Word Inverse frequency in corpus}}$$

Estimating Substring Semantic Information using TF-IDF



(b) Bag-of-Substrings model.

$$I(x_{i,j}, Sem(x)) = \underbrace{F(x_{i,j}, \mathbb{X}^p)}_{\text{Substring frequency in Paraphrases}} \underbrace{\log \frac{|\mathcal{D}|}{|x' : x_{i,j} \text{ is a substring of} x'|}}_{\text{Substring inverse frequency in corpus}}$$

Substring frequency in Paraphrases

Substring inverse frequency in corpus

Parsing by Max. Semantic Information

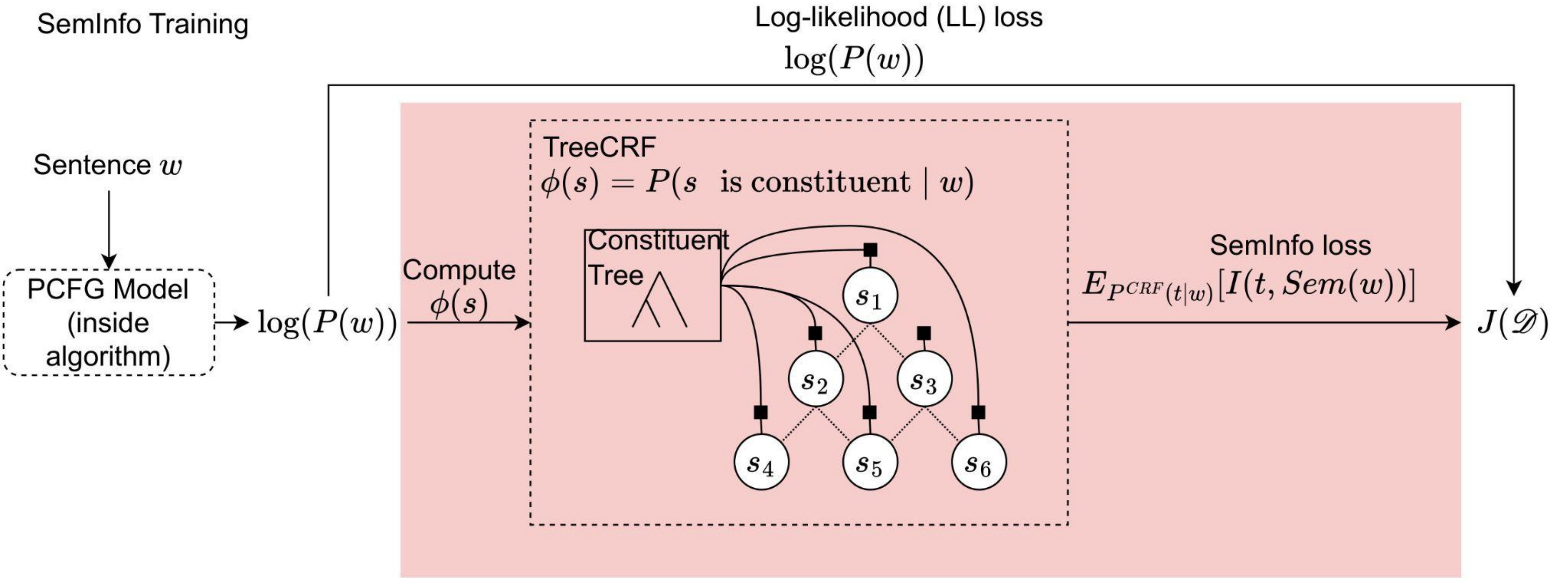
Estimating Substring Semantic Information $I(t, Sem(x))$ via paraphrasing

- Paraphrasing exposes semantic-driven word collocations
- Estimating substring semantic information using bag-of-substrings model and TF-IDF

Reinforcement learning to encourage high SemInfo predictions

- Training a parser by maximizing $I(t, Sem(x))$
- Enforcing Tree-constraint through RL

Learning Parsers by Max. SemInfo Training



SemInfo Maximization Results in Significant Accuracy Improvement

- SemInfo-trained PCFG parsers outperform LL-trained counterparts by a large margin
- The improvement is statistically significant in 17/20 test configurations

	English		Chinese		French		German	
	SemInfo (Ours)	LL	SemInfo	LL	SemInfo	LL	SemInfo	LL
CPCFG	$65.74_{\pm 0.81}$	$53.75_{\pm 0.81}$	$50.39_{\pm 0.87}$	$51.45_{\pm 0.49}$	$52.15_{\pm 0.75}$	$47.50_{\pm 0.41}$	$49.80_{\pm 0.31}$	$45.64_{\pm 0.73}$
NPCFG	$64.45_{\pm 1.13}$	$50.96_{\pm 1.82}$	$53.30_{\pm 0.42}$	$42.12_{\pm 3.07}$	$52.36_{\pm 0.62}$	$47.95_{\pm 0.09}$	$50.74_{\pm 0.28}$	$45.85_{\pm 0.63}$
SCPCFG	$67.27_{\pm 1.08}$	$49.42_{\pm 2.42}$	$51.76_{\pm 0.54}$	$46.20_{\pm 3.65}$	$52.79_{\pm 0.80}$	$45.03_{\pm 0.42}$	$47.97_{\pm 0.76}$	$45.50_{\pm 0.71}$
SNPCFG	$67.15_{\pm 0.62}$	$58.19_{\pm 1.13}$	$51.55_{\pm 0.82}$	$43.79_{\pm 0.39}$	$55.21_{\pm 0.47}$	$49.64_{\pm 0.91}$	$49.65_{\pm 0.29}$	$40.51_{\pm 1.26}$
TNPCFG	$66.55_{\pm 0.96}$	$53.37_{\pm 4.28}$	$51.79_{\pm 0.83}$	$45.14_{\pm 3.05}$	$54.11_{\pm 0.66}$	$39.97_{\pm 4.10}$	$49.26_{\pm 0.64}$	$44.94_{\pm 1.34}$
Average Δ	+13.09		+6.02		+7.31		+4.92	

PCFG Mitigates SemInfo Estimation Noise

- SemInfo-trained PCFG have higher accuracy than either SemInfo-only method and PCFG-only methods.
- SemInfo-trained PCFG benefits from even highly noisy paraphrasing model (qwen-0.5b and llama-1b)

	Paraphrasing Model Variations						
	Large Models			Medium Models		Small Models	
	gpt35	gpt4o	gpt4omini	llama3.2-3b	qwen2.5-3b	llama3.2 1b	qwen2.5-0.5b
SemInfo-NPCFG	66.85±0.25	65.19±0.54	64.45±1.13	63.78±0.55	63.58±0.13	63.10±0.70	59.01±0.24
SemInfo-MTD	55.56	59.45	58.28	55.17	55.03	48.5	43.3
LL-NPCFG	50.96±1.82						
Right Branching	38.4						

From hindsight

- Is language modeling (maximizing sentence likelihood, LL) sufficient to induce syntactic structure?
    - Partly, as LL maximization $\Longrightarrow$ a reasonable parser
    - **The semantic information shapes the structure**, as SemInfo is highly effective in inducing the structure
- Why structural analysis?
    - **Structure prior (PCFG) provides a denoising effect to semantic analysis**
    - Syntactic structure encodes information beyond semantic information (e.g., speaker intention)

Contribution

- Applied paraphrasing to **expose fine-grained latent semantics as paraphrase clusters in textual space**
- A novel training objective for unsupervised parsing revealing that **communication messages (semantics) shapes natural languages**

SemInfo is a better objective than LL:
SemInfo Ranks linguistically-correct trees better

- High coefficient → Ranking linguistically-correct trees appropriately
- SemInfo ranks linguistically-correct trees better than LL
  - Training on SemInfo approximates directly training on parsing accuracy.

	SemInfo-SF1^i	LL-SF1^i	SemInfo-LL
CPCFG	0.6518	0.0223	0.0196
NPCFG	0.6347	-0.0074	-0.0045
SCPCFG	0.6431	-0.0013	0.0505
SNPCFG	0.9289	0.0102	0.0182
TNPCFG	0.6449	0.1077	0.1426

Figure.1 Coefficient of SemInfo/LL – Sentence-level Parsing Accuracy correlation

SemInfo is a better objective than LL:
SemInfo Better Distinguishes Good Parsers from Bad ones

- High correlation → Ranking better parsers higher
- SemInfo correctly rank parsers throughout training, while LL correctly ranks parsers only at the early stage.

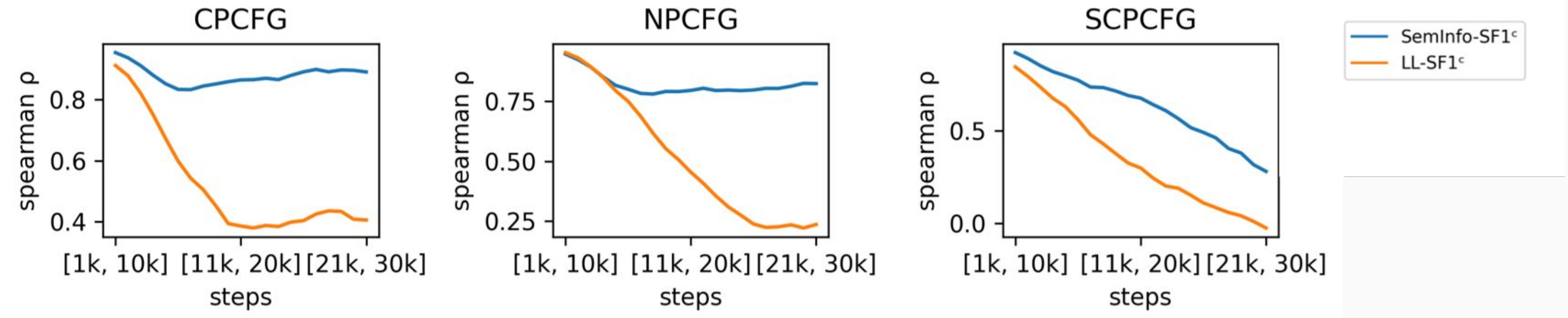


Figure.2 Change in SemInfo/LL – Parsing Accuracy correlation coefficient throughout training