Improving Unsupervised Constituency Parsing via Maximizing Semantic Information

Junjie Chen [1] Xiangheng He [2] Yusuke Miyao [1] Danushka Bollegala [3]

[1]The University of Tokyo [2]Imperial College London [3]University of Liverpool

Research Question

Can we predict the constituent structure by searching for a structure maximizing semantic information?

Motivation

- **Linguistically-defined constituent phrases often correspond to semantic concepts.**
- Constituent structures aid natural language understanding [2, 3].
- Constituent phrases are resilient against semantic-preserving perturbations [1].

} Constituents carry Semantic Information

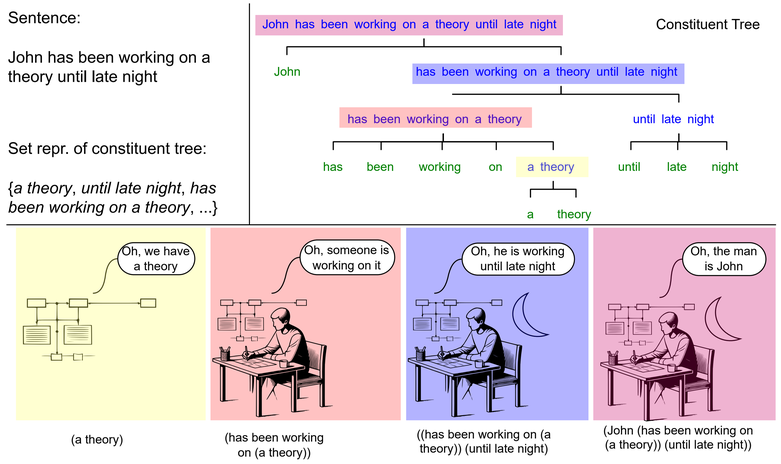

Figure 1. Correspondence between constituent phrases and semantic concepts.

Findings: Complementary Benefit of SemInfo and PCFG-induced Bias

SemInfo and PCFG-induced bias are complementary toward accurate unsupervised parsing.

- Three classes of paraphrasing models represent three levels of paraphrasing noises
- SemInfo improves PCFG parsing accuracy (SemInfo-NPCFG vs. LL-NPCFG)
- PCFG model improves parsing robustness (SemInfo-NPCFG vs. SemInfo-MaxTreeDecoding)

	Paraphrasing Model						
	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-MaxTreeDecoding	55.56	59.45	58.28	55.17	55.03	48.5	43.3
LL-NPCFG	50.96±1.82						
Right Branching	38.4						

Table 1. Parsing accuracy of parsers trained using SemInfo estimated from seven paraphrasing models

Reference

[1] A. Carnie. *Syntax: a generative introduction*. Introducing linguistics. Blackwell Pub, Malden, MA, 2nd ed edition, 2007.

[2] Q. He, H. Wang, and Y. Zhang. Enhancing generalization in natural language inference by syntax. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4973–4978, Online, Nov. 2020. Association for Computational Linguistics.

[3] P. Xie and E. Xing. A constituent-centric neural architecture for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1405–1414, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Method: Estimating Tree-Semantic Information (SemInfo)

Substring-Semantic Information

We propose to estimate substring-semantic information $I(w_{(i,j)}, Sem(w))$ via the following steps

- introducing a bag-of-substrings representation of sentence semantics based on a paraphrasing model.
- adapting the PWI metric in the bag-of-words model to estimate the substring-semantic information

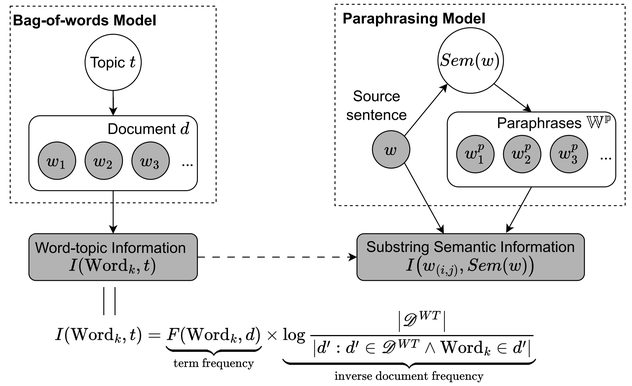

Figure 2. Structural parallelism between our bag-of-substrings model and the traditional bag-of-words model.

$$I(w_{(i,j)}, Sem(w)) = \underbrace{F(w_{(i,j)}, \mathbb{W}^p)}_{\text{substring frequency}} \times \log \underbrace{\frac{|\mathcal{D}^{SS}|}{|w' : w' \in \mathcal{D}^{SS} \wedge w_{(i,j)} \text{ is a substring of } w'|}}_{\text{inverse corpus frequency}}$$

SemInfo: Tree-Semantic Information

We calculate SemInfo $I(t, Sem(w))$ as the cumulative substring-semantic information associated with the tree t.

$$I(t, Sem(w)) = \sum_{w_{(i,j)} \in t} I(w_{(i,j)}, Sem(w))$$

Method: Training PCFG Parsers using SemInfo Maximization

We train a PCFG model by maximizing $I(t, Sem(w))$ through a TreeCRF model.

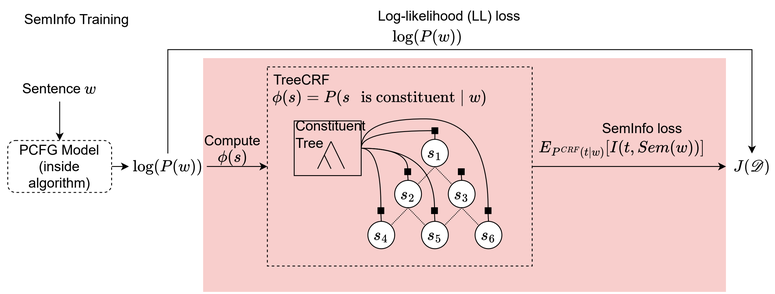

Figure 3. Proposed Training pipeline

Result: SemInfo Training Significantly Improves Parsing Accuracy

SemInfo-trained parsers **significantly** outperform LL-trained counterparts in **17**/20 combinations

	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	
SemInfo-MaxTreeDecoding	58.28		49.03		52.03		50.82	
GPT4o-mini	36.16		11.82		30.01		33.56	

Table 2. Parsing accuracy (SF1^c) of SemInfo-trained PCFG parsers and other baseline parsers.

Analysis: SemInfo Strongly Correlates with Parsing Accuracy

Why the SemInfo maximization training improves over the LL maximization training?

Sentence-level Analysis: SemInfo Ranking Approximates Parsing Accuracy Ranking

This analysis compares the ranking of predicted trees by SemInfo/LL and that by instance-level parsing accuracy SF1^i. High coefficient $\Longrightarrow$ SemInfo/LL ranking of constituent trees approximates that by parsing accuracy

- SemInfo has a strong correlation with parsing accuracy
- LL has negligible correlation with 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

Table 3. Spearman correlation coefficient between SemInfo/LL and SF1^i.

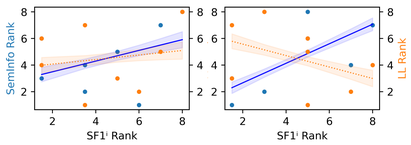

Figure 4. Scatter plot of (SemInfo, LL, SF1^i) pairs.

Corpus-level Analysis: SemInfo Consistently Ranks PCFG Parsers throughout Training

This analysis compares the ranking of parsers by SemInfo/LL and that by parsing accuracy SF1^c. High coefficient $\Longrightarrow$ SemInfo/LL can be applied to training PCFG parsers

- SemInfo maintains a consistently strong correlation with SF1^{c}throughout training.
- LL has a strong correlation with SF1^c at the early training stage, but the strength quickly degrades.

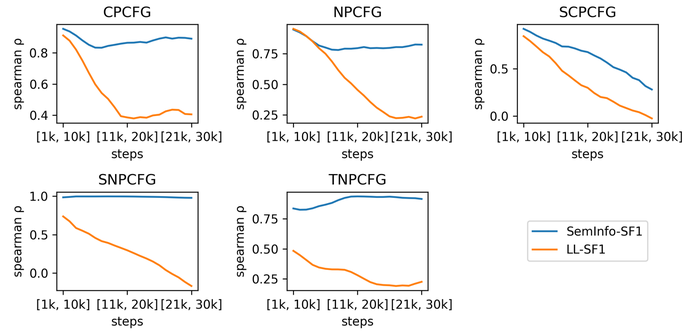

Figure 5. Change of corpus-level correlation throughout training

christopher@is.s.u-tokyo.ac.jp