Probing Syntactic Dependencies with Conditional Mutual Information and Grammatical Constraints

Anonymous ACL submission

Abstract

Unsupervised dependency parsing is a fundamental task in understanding syntactic dependency structures of natural language. Previous parameter-free methods for probing the dependency structure recover a non-trivial amount of dependencies by assuming a correlation between the syntactic dependency (a word-to-word relation) and bi-lexical dependence scores (a metric measuring one word's influence on the other word). However, these studies assume the correlation without verifying the existence of the correlation. Furthermore, previous studies failed to utilize grammatical constraints that are beneficial to parsing performance in grammar-based unsupervised parsing methods. In this paper, we investigate the correlation between the syntactic dependency and Conditional Mutual Information (CMI) scores, a bi-lexical statistical dependence metric. We propose *delta-energy*, an unbiased estimate of the CMI, and apply it to unsupervised dependency parsing. We further assist the parsing model with three grammatical constraints. We found the delta-energy score capable of effectively separating syntactic dependencies from non-dependencies. Our unsupervised parsing model outperforms baseline parameter-free probing models in parsing performance, excelling in recovering semantically-related dependencies. The ablation study shows that the three grammatical constraints contribute to the recovery of dependencies that are semantically related and that have strong Part-Of-Speech requirements.

1 Introduction

Syntactic dependency structures are important to downstream Natural Language Processing tasks, such as Information Extraction (Tian et al., 2021; Gamallo et al., 2012), Machine Translation (Bugliarello and Okazaki, 2020; Ma et al., 2020), and Question Answering (Lyu et al., 2021). However, training a supervised dependency parser requires expensive human-annotated dependency

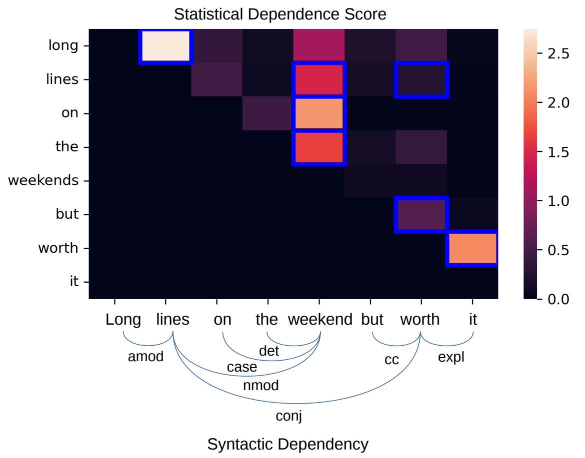


Figure 1: The correlation between syntactic dependencies and conditional mutual information scores. The lower part plots the dependency structure, and the upper part plots the conditional mutual information score. The blue box in the heatmap indicates a syntactic dependency between the corresponding words.

structures, which are only available for some languages and domains. Parameter-free probing methods (Hoover et al., 2021; Wu et al., 2020; Zhang and Hashimoto, 2021) directly extract the dependency structure from pre-trained language models without the structure annotation (a.k.a. unsupervised parsing). These methods predict the dependency structure by finding a set of dependencies that form a tree-shaped structure and that have maximum bi-lexical dependence scores.

Fig.1[1] illustrates the syntactic dependency structure and Conditional Mutual Information (CMI) scores, a bi-lexical statistical dependence metric. The syntactic dependency, represented as a word pair, encodes a word-to-word grammatical relation. For example, the dependency ("long", "line") with the label amod indicates that "long" serves as an

[1] Conditional mutual information is a symmetric score. We study the undirected syntactic dependency to put the conditional mutual information and other dependence scores on an equal footing.

adjective modifier to "line". On the other hand, the CMI score is a metric measuring one word's influence on the other word. The higher the CMI score, the stronger the influence. For example, the CMI score for the word pair ("long", "line") is high, indicating a strong bi-lexical dependence. The above example shows a correlation between syntactic dependencies and high CMI scores. Such correlation is the cornerstone of the parameter-free probing method.

Despite the cornerstone role of the correlation, the parameter-free probing method assumes the correlation without verifying it. Whether and how much their dependence scores separate the syntactic dependency from non-dependencies (i.e., word pairs that are not connected by a syntactic dependency) remains a question. Furthermore, the probing method failed to incorporate grammatical constraints that have been shown beneficial to parsing performance by grammar-based unsupervised parsing methods (Noji et al., 2016; Naseem et al., 2010; Xu et al., 2021).

In this paper, we present a study on the correlation between the syntactic dependency and the CMI score. Our contributions are three-fold:

1. We propose *delta-energy*, an unbiased estimate of the CMI score, and derive an unsupervised parsing model from the delta-energy score. We further enhance the parsing model with three grammatical constraints: a Part-Of-Speech (POS) constraint, an adjacent-connect constraint, and a function word head constraint.

2. We verify the correlation between the syntactic dependency and the delta-energy score and show that the delta-energy score is an effective metric for separating syntactic dependencies from non-dependencies.

3. We build a state-of-the-art parameter-free unsupervised parsing model that excels at recovering semantically-related dependencies. Ablation analysis shows that the grammatical constraint has a significant contribution to the recovery of dependencies that are semantically related and that tend to have strong POS requirements.

2 Background

2.1 Conditional Mutual Information

In this paper, we measure bi-lexical statistical dependence scores using Conditional Mutual Information (Eq. 1). Given a sentence $x = (x_1, ..., x_n)$, CMI measures one word's (X_i) influence on the other word's (X_j) distribution under side information c. The side information can include various types of information, contextual or grammatical. For example, the two words' context $x_{-ij} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_{j-1}, x_{j+1}, ..., x_n)$ can be the side information. CMI computes the expected log probability ratio between the joint probability $p(x_i, x_j|c)$ and the product of the marginal probabilities $p(x_i|c)p(x_j|c)$. CMI measures the distance between the joint distribution from the marginal product distribution. The higher the CMI score, the further the joint and the marginal product distribution are, the stronger the statistical dependence is between X_i and X_j.

$$I(X_i; X_j|c) := \mathop{\mathbb{E}}_{X_i X_j|c}\left[\log \frac{p(x_i, x_j|c)}{p(x_i|c)p(x_j|c)}\right] \quad (1)$$

2.2 Extracting Syntactic Dependency via Measuring Bi-Lexical Dependence

Hoover et al. (2021); Wu et al. (2020) measure the bi-lexical dependence score of word pairs (X_i, X_j) under the context x_{-ij}. They define the dependence score as the difference between the informed probability $p(x_j|x_i, x_{-ij})$ and the null probability $p(x_j|X_i = [\text{MASK}], x_{-ij})$ given by a masked language model. The informed probability measures the probability of x_j given the content of x_i, while the null probability measures the probability of x_j without knowing the content of x_i. The null probability can serve as an approximation to the marginal probability $p(x_j|x_{-ij})$ (Xu et al., 2020). After obtaining the dependence score for all word pairs, the probing method selects a dependency tree that maximizes the sum of the dependence scores using Maximum Spanning Tree algorithms (Prim, 1957; Eisner, 1997; Edmonds, 1967).

Hoover et al. (2021) uses a log ratio of the informed and the null probability (pmi, Eq. 2) as the dependence score. The pmi score is a single-point estimate of the CMI score, and the single-point estimation is well-known to have a high estimation variance. Moreover, the pmi score uses the null probability to approximate the marginal probability, which adds further bias to the estimation. The two issues of the pmi score make it an unreliable estimate of the CMI score. On the other hand, Wu et al. (2020) uses the Euclidean distance between the embedding generating the informed probability $e_{\text{informed\_prob}}$ and the embedding generating the null probability $e_{\text{null\_prob}}$ (Eq. 3). This method operates in the embedding space instead of the probabilistic space and is not directly comparable with our method.

$$I_{ij}^{pmi} := \log \frac{p(x_j|x_i, x_{-ij})}{p(x_j|X_i = \text{[MASK]}, x_{-ij})} \tag{2}$$

$$I_{ij}^{pert} := \|e_{\text{informed\_prob}} - e_{\text{null\_prob}}\|^2 \tag{3}$$

2.3 Sampling from Language Models

Monte Carlo estimation is the standard approach to estimating the CMI score reliably. Goyal et al. (2022) proposes a Metroplis-Hastings (MH) algorithm to sample from language models. Starting with a sentence x with arbitrary content on the word X_i and X_j, the MH method iteratively performs the following steps over the two words:

1. samples a proposal word x_i' from a proposal distribution $q(x_i'|x)$. In this case, the proposal distribution would be a mask language model distribution with X_i set to [MASK].

2. computes the acceptance probability based on the proposal probability and the target probability $p(x')$. Here, p is the language model distribution of choice and $x' = (x_1, ..., x_{i-1}, x_i', x_{i+1}, ..., x_n)$

3. accepts or rejects the proposal word according to the acceptance probability. If accepted, $x \leftarrow x'$. Otherwise, $x \leftarrow x$.

The (x_i, x_j) samples produced by the MH algorithm are guaranteed to converge to the target distribution $p(x_i, x_j|x_{-ij})$ as long as the target distribution is irreducible and aperiodic (Besag, 2004). Common language model distributions satisfy the irreducibility and aperiodicity conditions (Goyal et al., 2022).

Nonetheless, the convergence speed of the MH algorithm can be slow in practice. Multi-try MH algorithms (Martino, 2018) mitigate the slow convergence problem by independently proposing n samples and accepting the sample with the highest probability in the target distribution. This approach enables the multi-try MH algorithm to explore the high-probability region of the target distribution more efficiently than the original MH algorithm, leading to faster convergence.

3 Method

Our method consists of two stages: inducing CMI scores and decoding syntactic dependency structure from the CMI score. We incorporate the POS constraint in the induction stage and incorporate the adjacent-connect and the function word head constraint in the decoding stage.

3.1 Inducing CMI Scores

We define the bi-lexical dependence score as the CMI between two words X_i and X_j under side information c. The side information includes, mandatorily, the context x_{-ij} and, optionally, the POS constraint y_i and y_j for X_i and X_j. We compute the CMI with Eq. 4 and use a causal language model (CLM) distribution for p. Our experiments show that CLMs provide higher-quality samples than masked language models. Eq. 4 is an equivalent form of Eq. 1 that is more suitable for the sampling-based estimation of the CMI score. The first term is the expected probability of samples from the joint distribution $X_iX_j|c$, whereas the second term is the expected probability of samples from the marginal product distribution $X_i|c \otimes X_j|c$.

$$\begin{aligned} I_{ij}(x) = & \mathop{\mathbb{E}}_{(x_i,x_j)\sim X_iX_j|c} [\log p(x_i, x_j, c)] \\ & - \mathop{\mathbb{E}}_{(x_i',x_j')\sim X_i|c\otimes X_j|c} [\log p(x_i', x_j', c)] \end{aligned} \tag{4}$$

3.1.1 Sampling with the POS Constraint

Directly applying the MH algorithm to the CLM could not incorporate the POS constraint into the CMI score because the CLM predicts the next word based on its preceding context without considering any POS constraint. We incorporate the POS constraint by applying a mask over the CLM's output distribution. Rijkhoff (2007) points out that POS is a set of words with the same grammatical properties. By the definition, we can impose a POS constraint y_i for the word X_i by masking out words that do not have the grammatical property specified by the POS. We translate the idea into Eq.5. Eq.5 defines the conditional distribution $p(x_i|x_j, x_{-ij}, y_i)$ by renormalizing the CLM probability of a word $p(x_i|x_j, x_{-ij})$ with the total probability of all words satisfying the POS constraint $\sum_{X_i} p(x_i|x_j, x_{-ij})\mathbb{1}_{(Y(X_i)=y_i)}$

$$p(x_i|x_j, x_{-ij}, y_i) := \frac{p(x_i|x_j, x_{-ij})\mathbb{1}_{(Y(X_i)=y_i)}}{\sum_{X_i} p(x_i|x_j, x_{-ij})\mathbb{1}_{(Y(X_i)=y_i)}} \tag{5}$$

3.1.2 Estimating with the POS constraint

Estimating the CMI requires computing the joint probability of the sentence (x_i, x_j, x_{-ij}) and the POS constraint (y_i, y_j). Unfortunately, one can not compute the joint probability straightforwardly as the CLM does not model the POS constraint. We propose *delta-energy* (Eq.6), an unbiased estimate of the CMI score, to overcome this issue. Compared to the CMI score, the delta-energy score eliminates the POS constraint y_i and y_j inside the expectation, enabling straightforward computation of the probability using CLM. We prove that the elimination is safe such that the delta-energy score is equivalent to the CMI score when the side information c contains only the POS constraint and the remaining context (Appendix A.1)

$$I_{ij}^{DE}(x) = \mathop{\mathbb{E}}_{(x_i,x_j)\sim X_iX_j|c}[\log p(x_i, x_j, x_{-ij})] - \mathop{\mathbb{E}}_{(x_i',x_j')\sim X_i|c\otimes X_j|c}\left[\log p(x_i', x_j', x_{-ij})\right] \tag{6}$$

3.2 Decoding Syntactic Dependency from Delta-Energy Scores

We apply Prim's algorithm (Prim, 1957) to decode an undirected dependency tree from the symmetric delta-energy score. We additionally apply two grammatical constraints at this stage: the adjacent-connect constraint and the function-word head constraint.

The adjacent-connect constraint is a default strategy when a word is not statistically dependent on the rest of the sentence (i.e., $\forall j, I_{ij}(x) \approx 0$). In that case, we default the word to be connected with its right neighbor, inspired by the high parsing performance of a trivial baseline that connects adjacent words (Klein and Manning, 2004). We set a threshold τ such that a word is automatically connected to its right neighbor if the accumulative delta-energy score between the word and the rest of the sentence is below τ.

The function-word head constraint (Noji et al., 2016) prevents function words from being a syntactic head in the decoded structure. In the context of undirected dependencies, the constraint prevents function words from having more than one connection to other words. We enforce the constraint by gradually decreasing delta-energy scores related to the function word that violates the constraint. As we will see in Section 4, the two constraints are effective in improving the parsing performance.

4 Experiment

4.1 Experiment Setup

We use three datasets for experiments: EWT-10, WSJ-10 (Klein and Manning, 2004), and PUD. Among the three, the EWT-10 and the WSJ-10 dataset contain sentences shorter than 10 words (excluding punctuations) from the English Web Treebank (Bies, Ann et al., 2012) and the Penn Treebank (Marcus, Mitchell P. et al., 1999) respectively. The main reason for using the EWT-10 and the WSJ-10 datasets is the high computational cost of the delta-energy estimation. For example, running the delta-energy estimation on the development section of EWT-10 takes 48 GPU hours on a single A100 GPU. In addition, the WSJ-10 dataset is widely used for unsupervised dependency parsing (Klein and Manning, 2004; Cohen and Smith, 2009). The PUD dataset contains the full English section of the Parallel Universal Dependency treebank (Zeman et al., 2018). The EWT-10 and the PUD dataset contain dependencies annotated in the universal dependency format (Nivre et al., 2020) while the WSJ-10 contains dependencies annotated in the Stanford dependency format (de Marneffe and Manning, 2008). We use the development section of the EWT-10 dataset (i.e., EWT-DEV-10) to analyze the correlation between the syntactic dependency and the delta-energy score and to evaluate the parsing model derived from the delta-energy score. We use the test section of the EWT-10 dataset, the WSJ-10 dataset, and the PUD dataset to evaluate the models' parsing performance on universal dependencies, on Stanford dependencies, and on long sentences, respectively. The parsing performance is measured in Unlabelled Undirected Attachment Score (UUAS) (Nivre and Fang, 2017) due to the symmetricity of the delta-energy score. We compute the UUAS score for syntactic dependencies that connect actual words for all experiments (i.e., we exclude the root dependency from the evaluation).

We use the `bert-large-cased` model (Devlin et al., 2019) for the proposal distribution and the `gpt2-large` model (Radford et al., 2019) for the target distribution when sampling for the EWT-10 and the WSJ-10 datasets. For the PUD dataset, we

Dependence Scores	P-Value (↓)	Cohen's d (↑)
delta-energy	0.00E+00	1.14
pmi	2.38E-214	0.69
perturbed-masking	0.00E+00	1.11

Table 1: P-value and Cohen's d value for the three scores in separating the syntactic dependency from the non-dependency. The p-value indicates whether the score can separate the two dependencies, and the d value indicates the separation effect. A low p-value and a high d value indicate a good separation score.

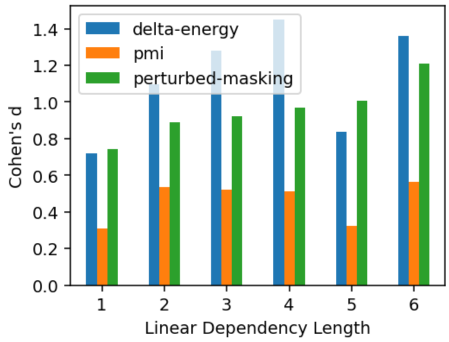


Figure 2: Cohen's d value by linear dependency lengths

alternatively use the `opt-125m` model (Zhang et al., 2022) for the target distribution to speed up the sampling process. We take one sample for every 12 sampling steps to avoid correlation between subsequent samples. In total, we collect 128 samples for every word pair. We limit words that can be sampled from the `bert-large-cased` model to the vocabulary of the bert model to reduce the computational and implementation complexity. We use the POS tag provided in the dataset for implementing the POS constraint and the function word head constraint. We run each experiment once because our method is parameter-free and also because of the high computational cost.

We use three baselines for analyses: the pmi baseline (Hoover et al., 2021), the perturbed-masking baseline [2] (Wu et al., 2020), and the adjacent-connect baseline (Klein and Manning, 2004).

4.2 Correlation between Syntactic Dependencies and Dependence Scores

The core question we seek to answer in this paper is: whether and to what degree does the syntactic dependency correlate with bi-lexical dependence scores? We answer this question by studying whether and how much the dependence score can separate the syntactic dependency from the non-dependency. We compare the delta-energy score with the dependence score derived from the pmi and the score derived from the perturbed-masking baseline. The experiment shows that the delta-energy score can separate and separates the syntactic and the non-dependency well.

The first column in Table 1 shows the p-value for the t-test with a null hypothesis that the syntactic dependency and the non-dependency have the same mean dependence score. All three scores have a p-value of 0 or close to 0, suggesting that all dependence scores can separate the syntactic and the non-dependency as two statistical populations.

The second column shows Cohen's d value, which measures the separation effect of the dependence score. The delta-energy score has the highest d value of 1.14 [3] among the three dependence scores. The perturbed-masking score has a medium d value of 1.11, and the pmi score has the lowest d value of 0.69. This result indicates that the delta-energy score is the best score for separating the dependency and the non-dependency group. The high d value suggests that the delta-energy model could perform better than the two baseline models in parsing performance, as we will see in the next section.

However, syntactic dependencies are not uniformly distributed across all dependency lengths. The syntactic dependency, on average, has shorter lengths than the non-dependency. The discrepancy creates a concern that the above analysis is not only measuring the separation of the syntactic dependency and the non-dependency but also the effect of the short and the long dependency. To counteract the concern, Fig. 2 breaks down the d value by the linear dependency length (i.e., the number of words between the word pair). The delta-energy score has the highest d value for most dependency lengths. The result reinforces the observation derived from Table 1 that the delta-energy score is the best score in separating the syntactic dependency from the non-dependency.

4.3 Parsing Performance of the Delta-Energy Model

Table 2 shows the parsing performance of the delta-energy model and the baseline parsing models on

[2] we use the released code for the experiment but corrected an implementation bug. See Appendix.A.3

[3] A d value of 0.5 indicates a medium effect, 0.8 a large effect, and 1.2 a very large effect (Sawilowsky, 2009)

Dependence Score	UUAS
delta-energy	**0.631**
pmi	0.559
perturbed-masking	0.586
adjacent-connect	0.497

Table 2: UUAS scores of the delta-energy and the baseline models on the EWT-DEV-10 dataset

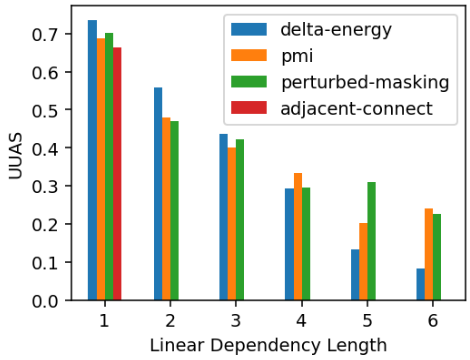


Figure 3: UUAS scores for syntactic dependencies of different lengths on the EWT-DEV-10 dataset

the EWT-DEV-10 dataset. The delta-energy model performs the best, leading the perturbed-masking model (the second-best model) by 0.046. The delta-energy, the perturbed-masking, and the pmi model outperform the adjacent-connect baseline by a large margin. The result confirms that the delta-energy score is the best score for separating the syntactic dependency from the non-dependency. The high performance of the delta-energy, the perturbed-masking, and the pmi model indicates that one can recover a non-trivial amount of syntactic dependencies by measuring the bi-lexical dependence score.

Fig. 3 plots the UUAS scores of the delta-energy model and the baseline models for recovering the syntactic dependency of different lengths. The delta-energy model performs the best for the syntactic dependency with lengths up to 3, performs similarly to the perturbed-masking model for the syntactic dependency with a length of 4, and performs the worst for the syntactic dependency with lengths 5 and 6. The result reveals the source of the delta-energy model's improvement: the short-length dependency, which makes up the majority of the syntactic dependency. For example, the EWT-DEV-10 dataset has 484 dependencies of lengths greater or equal to 4 while containing 4036 dependencies of lengths less than 4.

Fig. 4 shows the UUAS scores for relations

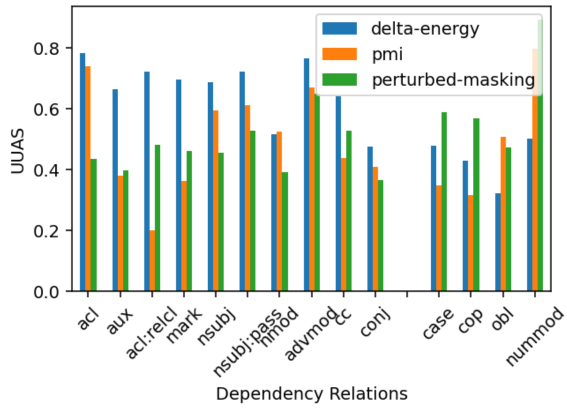


Figure 4: UUAS of relations where the performance difference between the delta-energy and the perturbed-masking model is greater than 0.1. The left part plots the relations where the delta-energy model performs better, and the right part plots the relations where the perturbed-masking model performs better.

where the performance difference between the delta-energy and the perturbed-masking model is more than 0.1. The delta-energy model outperforms the perturbed-masking model in recovering the semantically-related dependencies while underperforming in recovering functionally-related dependencies. The result indicates that the delta-energy model is more sensitive to semantically-related dependencies than functionally-related dependencies.

4.4 Ablation Study

model	UUAS (% loss)	Precision (% loss)
+UPOS, +ADJC, +FNWH	0.6313 (0.00%)	0.6332 (0.00%)
+UPOS, +ADJC, -FNWH	0.6064 (-3.94%)	0.6064 (-4.23%)
+UPOS, -ADJC, +FNWH	0.5518 (-12.59%)	0.5934 (-6.29%)
+UPOS, -ADJC, -FNWH	0.5319 (-15.75%)	0.5712 (-9.79%)
-UPOS, +ADJC, +FNWH	0.5937 (-5.96%)	0.5966 (-5.78%)
-UPOS, +ADJC, -FNWH	0.5544 (-12.18%)	0.5544 (-12.44%)
-UPOS, -ADJC, +FNWH	0.5403 (-14.41%)	0.58 (-8.40%)
-UPOS, -ADJC, -FNWH	0.4992 (-20.93%)	0.535 (-15.51%)

Table 3: Ablation analysis for the delta-energy model. +UPOS indicates the use of the POS constraint, +ADJC indicates the adjacent-connect constraint, and +FNWH indicates the function word head constraint.

Table 3 presents an ablation study for the three grammatical constraints. The UPOS, ADJC, and FNWH represent the POS, the adjacent-connect, and the function word head constraint, respectively. The +/- sign indicates whether the model uses the constraint. The Table shows that removing the POS or the FNWH constraint equally decreases the UUAS and the precision score. On the other hand, removing the ADJC constraint decreases the UUAS score more than the precision score. This is because, in some cases, the delta-energy score measures a 0 dependence score between one word and the rest of the sentence. The 0 dependence score creates an orphan problem in that the word is statistically disconnected from the rest of the sentence, resulting in an underprediction of the syntactic dependency. The ADJC constraint mitigates the orphan problem by forcibly connecting the word to its right neighbor. The ablation study with the UUAS score indicates that all grammatical constraints benefit the parsing performance and that the adjacent-connect constraint is important in resolving the orphan problem.

Fig. 5 analyzes which relation the grammatical constraint helps the most. The figures plot the UUAS for the relation where removing the respective grammatical constraint causes more than 0.1 loss in UUAS. Fig. 5a shows that the ADJC constraint improves performance for a wide range of dependencies. Fig. 5b shows that the POS constraint improves performance for dependencies with strong POS requirements. For example, the conj relation requires two words to have the same POS tag. The parataxis relation also includes cases where the two words have the same POS tag (Nivre et al., 2020). Fig. 5c shows that the FNWH constraint improves performance for semantically-related dependencies. The `nsubj`, `obl`, and `advcl` dependencies connect words with their semantic arguments. The `nmod` and `acl` relation connect words with their modifiers. These dependencies contribute to the semantics of the sentence. The result suggests that grammatical constraints are important for decoding syntactic dependencies from language models.

4.5 Comparison with State-of-the-Arts in Unsupervised Parsing

Type	Models	EWT-TEST-10	WSJ-10	PUD
Parameter-free Probing Models	delta-energy	**0.615**	**0.592**	**0.525**
	perturbed-masking	0.591	0.584	0.507
	mlmbias	0.352	0.586	0.495
Grammar-based Models	dmv	0.611	0.597	0.484
	lcdmv	0.659	0.614	0.554

Table 4: Comparison of the delta-energy model with unsupervised parsing models. The best score for the parameter-free probing models is in bold.

Table 4 compares the delta-energy model with two parameter-free probing models (perturbed-masking and mlmbias (Zhang and Hashimoto, 2021)) and two parametric grammar-based models (dmv (Klein and Manning, 2004) and lcdmv (Noji et al., 2016)) [4]. The dmv model has the same grammatical constraint as the delta-energy model, while the lcdmv model has an additional constraint that the sentence can not have a deep recursive center-embedding structure (Noji et al., 2016). Table 4 shows that the delta-energy model performs the best among the parameter-free probing models. Compared to the dmv model, the delta-energy model performs better on the PUD dataset and performs similarly on the EWT-TEST-10 and the WSJ-10 datasets. The better performance on the PUD dataset highlights the strength of the delta-energy model in comparison with the dmv model, a grammar-based model using the same grammatical constraint. Nonetheless, the delta-energy model falls behind the lcdmv model because the lcdmv model has access to additional grammatical constraints. The result again reinforces the importance of the grammatical constraint in recovering syntactic dependencies from language models.

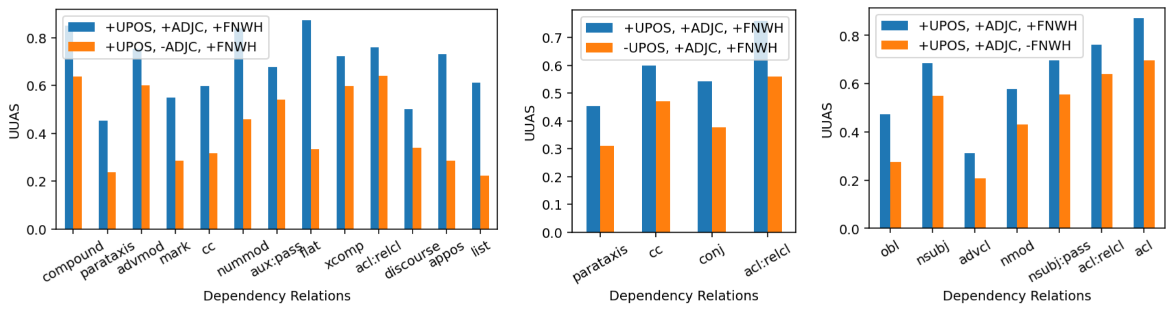
5 Related Works

5.1 Parameter-free Probing Methods

The pmi score (Hoover et al., 2021) measures the bi-lexical dependence score using the log-ratio between the informed and the null probability given by the BERT model. Besides the estimation problem mentioned in Section 2, the pmi score failed to utilize the POS information. In comparison, our method can utilize the POS information as a constraint and improves parsing performance with the information.

The perturbed-masking score (Wu et al., 2020) measures the bi-lexical dependence score using the Euclidean distance of the embedding that generates the informed and the null probability. Despite the simple approach, the perturbed-masking score performs well in recovering the syntactic dependency. However, the perturbed-masking score operates in the embedding space, making it difficult to establish a direct connection between the syntactic dependency and the language modeling objective. In contrast, our delta-energy score operates in the probabilistic space and, consequently, can establish a more direct connection with the language modeling objective. Furthermore, the perturbed masking score cannot utilize the POS information like the

[4] We use the code released by Noji et al. (2016) for the dmv and the lcdmv model



(a) adjacent-connect constraint (b) POS constraint (c) function word head constraint

Figure 5: Ablation analysis by dependency relations where removing the respective constraint causes more than 0.1 loss in UUAS score

pmi score, whereas our delta-energy score can utilize the POS information for better performance.

Zhang and Hashimoto (2021) measures the bi-lexical dependence score using their formulation of the "conditional mutual information". However, their formulation has two theoretical issues. Firstly, their formulation has an upper bound of 0, in contrast to the widely-known CMI, which is a strictly non-negative metric. Secondly, two statistically independent variables can obtain the maximum value under their formulation. The two issues disqualify their formulation as a valid dependence score. We present the proof in Appendix A.2.

5.2 Parametric Grammar-based Methods

The Grammar-based parametric method (Klein and Manning, 2004; Noji et al., 2016) induces grammar by maximizing the likelihood of observed sentences. While the method can theoretically avoid the data availability problem of lacking dependency annotations, most studies assume the POS information (Han et al., 2020). Effectively, the grammar-based method is still constrained by the availability of the POS information. On the other hand, our method can utilize the POS information as supplementary information. Moreover, the grammar-based method requires a special initialization (Klein and Manning, 2004; Yang et al., 2020) or grammatical constraints (Noji et al., 2016) to induce grammar successfully. As shown in Sec.4, our method can benefit from the constraint but does not require the constraint to extract the dependency properly.

6 Conclusions

In this paper, we studied the correlation between syntactic dependencies and CMI scores derived from causal language models and the application of the CMI score on unsupervised parsing. We proposed delta-energy, an unbiased estimate of the CMI score that allows the incorporation of POS constraints. We verified that syntactically connected words are more statistically dependent under causal language model distributions. The delta-energy score is the best metric for separating syntactic dependencies from non-dependencies. We found that the unsupervised parsing model induced by the delta-energy score outperforms baseline models by a large margin. The delta-energy model outperforms baseline models in recovering semantically-related dependencies but underperforms in recovering functionally-related dependencies. Our ablation study shows that the POS, the adjacent-connect, and the function word head constraint benefit the parsing performance. The POS constraint contributes to the recovery of dependencies with strong POS requirements. The adjacent-connect constraint boosts the performance in recovering a wide range of dependencies. The function word head constraint significantly contributes to recovering semantically related dependencies. The result indicates the importance of grammatical constraints in extracting syntactic dependencies from language models. The delta-energy model performs strongly against state-of-the-art parameter-free probing models and matches the performance of the grammar-based parametric model using similar grammatical constraints.

7 Limitations

A concern we have is the high computational cost of the MH algorithm. At every sampling step, the MH algorithm has to evaluate the probability of the sentence $x' = (x_1, ..., x_{i-1}, x'_i, x_{i+1}, ..., x_n)$ with the proposal sample x'_i. Since we collect 128 samples and take one sample for every 12 sampling steps, we have to evaluate the sentence probability 1536 times to estimate the CMI score for a word pair. The above is for the constant factor of the computational complexity. The total computational complexity for a sentence with n words is $O(n^5)$ considering that the CLM model has a computational complexity of $O(n^3)$ ($O(n^2)$ complexity for one pass through the transformer model and $O(n)$ steps to obtain the probability for each word in the sentence). The high computational cost prevents us from conducting a large-scale multilingual experiment for languages that have dependency annotations (Nivre et al., 2020) available.

References

Julian Besag. 2004. An introduction to markov chain monte carlo methods. In *Mathematical Foundations of Speech and Language Processing*, pages 247–270. Springer New York.

Bies, Ann, Mott, Justin, Warner, Colin, and Kulick, Seth. 2012. English web treebank.

Emanuele Bugliarello and Naoaki Okazaki. 2020. Enhancing machine translation with dependency-aware self-attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1618–1627, Online. Association for Computational Linguistics.

Shay Cohen and Noah A. Smith. 2009. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 74–82, Boulder, Colorado. Association for Computational Linguistics.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK. Coling 2008 Organizing Committee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jack Edmonds. 1967. Optimum branchings. *Journal of Research of the National Bureau of Standards Section B Mathematics and Mathematical Physics*, 71B(4):233.

Jason Eisner. 1997. Bilexical grammars and a cubic-time probabilistic parser. In *Proceedings of the Fifth International Workshop on Parsing Technologies*, pages 54–65, Boston/Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Pablo Gamallo, Marcos Garcia, and Santiago Fernández-Lanza. 2012. Dependency-based open information extraction. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, pages 10–18, Avignon, France. Association for Computational Linguistics.

Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. 2022. Exposing the implicit energy networks behind masked language models via metropolis–hastings. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Wenjuan Han, Yong Jiang, Hwee Tou Ng, and Kewei Tu. 2020. A survey of unsupervised dependency parsing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2522–2533, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jacob Louis Hoover, Wenyu Du, Alessandro Sordoni, and Timothy J. O'Donnell. 2021. Linguistic dependencies and statistical dependence. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2941–2963, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dan Klein and Christopher Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 478–485, Barcelona, Spain.

Chenyang Lyu, Lifeng Shang, Yvette Graham, Jennifer Foster, Xin Jiang, and Qun Liu. 2021. Improving unsupervised question answering via summarization-informed question generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4134–4148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nianzu Ma, Sahisnu Mazumder, Hao Wang, and Bing Liu. 2020. Entity-aware dependency-based deep graph attention network for comparative preference classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5782–5788, Online. Association for Computational Linguistics.

Marcus, Mitchell P., Santorini, Beatrice, Mary Ann Marcinkiewicz, and Taylor, Ann. 1999. Treebank-3.

Luca Martino. 2018. A review of multiple try MCMC algorithms for signal processing. *Digit. Signal Process.*, 75:134–152.

Tahira Naseem, Harr Chen, Regina Barzilay, and Mark Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1234–1244, Cambridge, MA. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Joakim Nivre and Chiao-Ting Fang. 2017. Universal Dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 86–95, Gothenburg, Sweden. Association for Computational Linguistics.

Hiroshi Noji, Yusuke Miyao, and Mark Johnson. 2016. Using left-corner parsing to encode universal structural constraints in grammar induction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 33–43, Austin, Texas. Association for Computational Linguistics.

R. C. Prim. 1957. Shortest connection networks and some generalizations. *Bell System Technical Journal*, 36(6):1389–1401.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jan Rijkhoff. 2007. Word classes. *Language and Linguistics Compass*, 1(6):709–726.

Shlomo S. Sawilowsky. 2009. New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8(2):597–599.

Yuanhe Tian, Guimin Chen, Yan Song, and Xiang Wan. 2021. Dependency-driven relation extraction with attentive graph convolutional networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4458–4471, Online. Association for Computational Linguistics.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhiyang Xu, Andrew Drozdov, Jay Yoon Lee, Tim O'Gorman, Subendhu Rongali, Dylan Finkbeiner, Shilpa Suresh, Mohit Iyyer, and Andrew McCallum. 2021. Improved latent tree induction with distant supervision via span constraints. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4818–4831, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Songlin Yang, Yong Jiang, Wenjuan Han, and Kewei Tu. 2020. Second-order unsupervised neural dependency parsing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3911–3924, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Tianyi Zhang and Tatsunori B. Hashimoto. 2021. On the inductive bias of masked language modeling: From statistical to syntactic dependencies. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5131–5146, Online. Association for Computational Linguistics.

A Appendix

A.1 Equivalence of Delta-Energy and CMI

Proposition 1. *I_{ij}^{DE} is equivalent with I_{ij} when the side information c contains only the POS information (y_i, y_j) and the remaining context x_{-ij}.*

We first look at the definition of CMI. Let $c = (x_{-ij}, y_i, y_j)$

$$
\begin{aligned}
& I_{ij}(x) && (7)\\
& := \mathbb{E}_{X_i X_j|c}\left[\log \frac{p(x_i, x_j|c)}{p(x_i|c)p(x_j|c)}\right] && (8)\\
& = \mathbb{E}_{X_i X_j|c}\left[\log \frac{p(x_i, x_j, c)}{p(x_i|c)p(x_j, c)}\right] && (9)\\
& = \mathop{\mathbb{E}}_{\substack{(x_i,x_j)\sim X_i X_j|c \\ (x_i',x_j')\sim X_i|c \otimes X_j|c}}\left[\log \frac{p(x_i, x_j, c)}{p(x_i'|c)p(x_j', c)}\right] && (10)\\
& = \mathop{\mathbb{E}}_{\substack{(x_i,x_j)\sim X_i X_j|c \\ (x_i',x_j')\sim X_i|c \otimes X_j|c}}\left[\log \frac{p(x_i, x_j, c)}{p(x_i', x_j', c)}\right] && (11)\\
& = \mathop{\mathbb{E}}_{\substack{(x_i,x_j)\sim X_i X_j|c \\ (x_i',x_j')\sim X_i|c \otimes X_j|c}}\left[\log \frac{p(x_i, x_j, x_{-ij})\mathbb{1}_{Y(x_i,x_j)=y_i,y_j}}{p(x_i', x_j', x_{-ij})\mathbb{1}_{Y(x_i',x_j')=y_i,y_j}}\right] && (12)\\
& = \mathop{\mathbb{E}}_{\substack{(x_i,x_j)\sim X_i X_j|c \\ (x_i',x_j')\sim X_i|c \otimes X_j|c}}\left[\log \frac{p(x_i, x_j, x_{-ij})}{p(x_i', x_j', x_{-ij})}\right] && (13)\\
& = I_{ij}^{DE}(x) && (14)
\end{aligned}
$$

The key to the proof lies in that the samples always satisfy the condition $Y(x_i, x_j) = y_i, y_j$ and $Y(x_i', x_j') = y_i, y_j$. Consequently, the indicator function will always return 1 and enables us to safely remove the POS information inside the expectation.

A.2 Theoretical Issues of Zhang and Hashimoto (2021)

They proposed a formulation of "conditional mutual information" (Eq.15)

$$
\begin{aligned}
I_{ij}^{ZH}(x) = \mathop{\mathbb{E}}_{X_i X_j|x_{-ij}} \Big[& \log p(x_i|x_j, x_{-ij}) \\
& - \log \mathop{\mathbb{E}}_{X_j|x_i,x_{-ij}} p(x_i|x_j, x_{-ij}) \Big] \qquad (15)
\end{aligned}
$$

We prove the following propositions

Proposition 2. *The upper bound of I_{ij}^{ZH} is 0.*

Proof.

$$
\begin{aligned}
(15) &= \mathop{\mathbb{E}}_{X_i|x_{-ij}} \Big[\mathop{\mathbb{E}}_{X_j|x_i,x_{-ij}} \log p(x_i|x_j, x_{-ij}) \\
&\qquad - \log \mathop{\mathbb{E}}_{X_j|x_i,x_{-ij}} p(x_i|x_j, x_{-ij}) \Big] && (16)\\
&\leq \mathop{\mathbb{E}}_{X_i|x_{-ij}} \Big[\mathop{\mathbb{E}}_{X_j|x_i,x_{-ij}} \Big[\log p(x_i|x_j, x_{-ij}) \\
&\qquad - \mathop{\mathbb{E}}_{X_j|x_i,x_{-ij}} \log p(x_i|x_j, x_{-ij}) \Big] \Big] && (17)\\
&= 0 && (18)
\end{aligned}
$$

□

Proposition 3. *Two statistically independent can reach the maximum value of 0 under I_{ij}^{ZH}.*

Proof. Let the two random variables be defined over a two-value set $X_i, X_j = \{0, 1\}$. Each value has a probability of 0.5. Consequently, we have the joint and the marginal probability as shown in the following table.

X_i		0	1
X_j	Prob	0.5	0.5
0	0.5	0.25	0.25
1	0.5	0.25	0.25

$$
\begin{aligned}
I^{ZH}(X_i; X_j) = (2 * 0.5)\Big[& (0.5 * 2) \log 0.5 && (19)\\
& - \log(0.5 * 2 * 0.5) \Big] && (20)\\
= 0 & && (21)
\end{aligned}
$$

□

A.3 Implementation bug in Wu et al. (2020)

Datasets	Released	Corrected
EWT-DEV-10	0.626	0.586
EWT-TEST-10	0.581	0.591
WSJ-10	0.572	0.584
PUD	0.501	0.507

Table 5: UUAS of the released implementation vs. our corrected implementation on four datasets

Wu et al. (2020) applies a softmax normalizing the dependence score I_{ij} based on all dependence scores $I_{i\cdot}$ related to the word X_i. However, we found that their implementation produces a normalized score with a value far greater than 1, which is impossible as the softmax function produces results in the range of $[0, 1]$. We fix the bug using the implementation provided in the scipy (Virtanen et al., 2020) package. Table 5 compares the performance of the released implementation and our corrected implementation. We see that, except for the EWT-DEV-10 dataset, the corrected implementation outperforms the released implementation. This suggests that our implementation does not artificially lower the method's performance in any way.