Language Model Based Unsupervised Dependency Parsing with Conditional Mutual Information and Grammatical Constraints

Junjie Chen
The University of Tokyo
junjiechen@g.ecc.u-tokyo.ac.jp

Xiangheng He
Imperial College London
x.he20@imperial.ac.uk

Yusuke Miyao
The University of Tokyo
yusuke@is.s.u-tokyo.ac.jp

Abstract

Previous methods based on Large Language Models (LLM) perform unsupervised dependency parsing by maximizing bi-lexical dependence scores. However, these previous methods adopt dependence scores that are difficult to interpret. Furthermore, these methods cannot incorporate grammatical constraints that previous grammar-based parsing research has shown beneficial to improving parsing performance. In this work, we apply Conditional Mutual Information (CMI), an interpretable metric, to measure the bi-lexical dependence and incorporate grammatical constraints into LLM-based unsupervised parsing. We incorporate Part-Of-Speech information as a grammatical constraint at the CMI estimation stage and integrate two additional grammatical constraints at the subsequent tree decoding stage. We find that the CMI score positively correlates with syntactic dependencies and has a stronger correlation with the syntactic dependency than baseline scores. Our experiment confirms the effectiveness and applicability of the proposed grammatical constraints across five languages and eight datasets. The CMI parsing model outperforms state-of-the-art LLM-based models and similarly constrained grammar-based models. Our analysis reveals that the CMI model is strong in retrieving dependency relations with rich lexical interactions but is weak in retrieving relations with sparse lexical interactions, indicating a potential limitation in CMI-based unsupervised parsing methods.

1 Introduction

Syntactic dependency structures provide important information 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, extracting the dependency structure using supervised methods requires expensive human-annotated dependency structures, which are only available for limited languages and domains. Large Language Model (LLM) based unsupervised parsing methods (Hoover et al., 2021; Wu et al., 2020; Zhang and Hashimoto, 2021) circumvent the problem by directly extracting the dependency structure from LLMs. These methods estimate bi-lexical dependence scores from the LLMs and identify the dependency structure as the tree with maximum bi-lexical dependence scores.

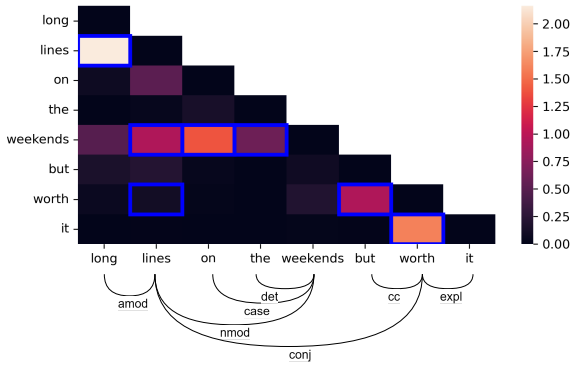

Figure 1: Illustration of the correlation between syntactic dependencies and bi-lexical dependence scores using CMI scores. The upper figure depicts the CMI scores, while the lower figure shows the dependency structure. Blue boxes in the figure indicate a syntactic dependency between the two corresponding words.

Figure 1 illustrates the motivation using our proposed bi-lexical dependence score. We can observe a positive correlation that syntactically dependent words (i.e., syntactic dependencies) tend to have a higher-than-average dependence score. With the positive correlation, we can factor the unsupervised parsing problem into two subproblems:(1) devising a dependence score that correlates well with the syntactic dependency and (2) performing a Maximum Spanning Tree (MST) decoding. The strength of the correlation between the dependence score and the syntactic dependency directly impacts the

We release our code in the github repository.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6355–6366

unsupervised parsing performance.

However, some LLM-based methods (Wu et al., 2020; Jian and Reddy, 2023) adopt dependence scores that lack a statistical interpretation of why some word pairs are more dependent than others. Other methods (Hoover et al., 2021; Zhang and Hashimoto, 2021) suffer from significant estimation issues. Additionally, no LLM-based method can incorporate grammatical constraints that previous grammar-based parsing research (Noji et al., 2016; Naseem et al., 2010; Xu et al., 2021) has shown beneficial to parsing performance.

In this paper, we apply Conditional Mutual Information (CMI) (Cover and Thomas, 1991), an interpretable metric, for measuring the bi-lexical dependence relation. CMI reveals the statistical correlation of two words under a given context, with a high CMI score indicating a strong correlation. We apply the Metropolis-Hasting sampling method (Hastings, 1970) to obtain an unbiased CMI estimate. We incorporate Part-Of-Speech (POS) information into the CMI estimation process as the POS constraint and further integrate two grammatical constraints (i.e., Adjacent-Connect (AC) constraint and Function Word UnHeading (FNUH) constraint) in the dependency tree decoding stage.

Our main contributions are three-fold. (1) We found a positive correlation that syntactically dependent words tend to have a high CMI (i.e., they are more likely to correlate). (2) We found the grammatical constraints effective in improving parsing performance and generally applicable across five languages. (3) We found that the CMI model performs strongly in dependency relations with rich lexical interactions while performing weakly in relations with sparse lexical interactions. Our study confirms the benefit of grammatical constraints in LLM-based unsupervised dependency parsing while suggesting a limit of CMI-based methods in the unsupervised parsing task.

2 Background

2.1 Bi-lexical Dependence Scores and CMI

Previous LLM-based methods (Wu et al., 2020; Hoover et al., 2021; Zhang and Hashimoto, 2021) use bi-lexical dependence scores as their basis for parsing. Given a sentence $x = (x_1, ..., x_n)$, the bi-lexical dependence score for (x_i, x_j) is computed as the distance between two events: $X_j|x_i, x_{-ij}$ and $X_j|x_{-ij}$ where $x_{-ij} := (x_1, ..., x_{i-1}, x_{i+1}, ..., x_{j-1}, x_{j+1}, ..., x_n)$. The former is the event of generating X_j with x_i and the context x_{-ij} while the latter is the event of generating X_j with only the context.

Conditional Mutual Information (CMI) measures the statistical correlation of x_i and x_j by one word's impact on the distribution of the other word. A high CMI score means changing one word will drastically change the distribution of the other word, indicating a strong bi-lexical dependence. Given a probability measure p, the CMI score is computed as the Kullback–Leibler divergence (a statistical distance) between the above two events (Equation 1).

$$I_{ij}^{cmi} := \mathbb{E}_{X_i X_j | x_{-ij}} \left[\log \frac{p(x_j|x_i, x_{-ij})}{p(x_j|x_{-ij})} \right] \quad (1)$$

2.2 MH Sampling from LLMs

Metroplis-Hasting (MH) sampling has been widely applied to collect random samples from target distributions where we can evaluate sample probability but cannot perform direct sampling (Singh et al., 2012; Miao et al., 2019). LLM distributions fall precisely into this category. Goyal et al. (2022) proposes an MH sampler to gather sentential samples using the LLM estimated sentence probability as the sample probability. The MH sampler performs iterative sampling for each token. At each step, the sampler performs the following procedure:

1. samples a proposal word $x_i^{k\prime}$ from a proposal distribution $q(X_i|x_{-i}^{k-1})$ at step k for the i-th word.

2. accepts the proposed sample by probability $\min(1, \frac{q(x_i^{k-1}|x_{-i}^{k-1})p(x_i^{k\prime}, x_{-i}^{k-1})}{q(x_i^{k\prime}|x_{-i}^{k-1})p(x^{k-1})})$. If accepted, $x_i^k \leftarrow x_i^{k\prime}$. Otherwise, $x_i^k \leftarrow x_i^{k-1}$.

The relative probability $\frac{p(x_i^{k\prime}, x_{-i}^{k-1})}{p(x^{k-1})}$ dictates that words resulting in higher sentence probability would be sampled more often than words resulting in lower sentence probability. The MH sampler returns all recorded $\{(x_1^k, ..., x_n^k)\}_k$ as random samples from the sentence distribution.

3 Related Works

Perturbed Masking Score Perturbed Masking (PM) (Wu et al., 2020) computes the bi-lexical dependence score as the Euclidean distance between two representations $e_j^{f_j^{\text{MASK}}(x)}$ (representing

the event $X_j|x_i, x_{-ij}$) and $e_j^{f_{ij}^{\text{MASK}}(x)}$ (representing the event $X_j|x_{-ij}$) (Equation 2). Here, the function $f_{ij}^{\text{MASK}}(x)$ is a masking function that sets the i-th and j-th token to the masked token and e_j^x is the masked LLM embedding for the j-th token when given x as input. However, the PM score is difficult to interpret because what the Euclidean distance means is unclear. Our CMI-based method provides higher interpretability, as a high CMI score indicates a strong statistical correlation.

$$I_{ij}^{pm} := \|e_j^{f_j^{\text{MASK}}(x)} - e_j^{f_{ij}^{\text{MASK}}(x)}\|_2 \tag{2}$$

Other CMI-based Scores Hoover et al. (2021) and Zhang and Hashimoto (2021) attempted unsupervised parsing with their own CMI implementation. Hoover et al. (2021) computes the CMI score as a point-wise estimate, using the original words (x_i, x_j) for estimation and replacing the marginal probability term $p(x_j|x_{-ij})$ with $p(x_j|X_i = [\text{MASK}], x_{-ij})$. The point-wise estimate suffers from high estimation variance, and replacing the marginal probability term introduces additional biases. The two issues explain the low parsing performance reported. Zhang and Hashimoto (2021) estimates the CMI score via Gibbs sampling and computes the CMI using their unique formulation. Their CMI formulation is theoretically ill-founded, as explained in Appendix A.2. The theoretical issue disqualifies Zhang and Hashimoto (2021)'s score as a valid dependence score and explains the low-performance figure shown in Table 7. Our method adheres closely to the CMI definition and provides a more reliable estimate using a Multi-Try MH sampler.

Grammar-based Methods Grammar-based unsupervised parsing is a parallel line of research to the LLM-based parsing method. The grammar-based methods induce a dependency grammar from plain text and perform parsing using the grammar. Previous grammar-based methods (Noji et al., 2016; Yang et al., 2020) achieve high parsing performance that no LLM-based method can match. However, these grammar-based methods require grammatical constraints or linguistic priors, such as locality bias (Smith and Eisner, 2006; Cohen and Smith, 2009; Klein and Manning, 2004), structural constraints (Noji et al., 2016), and grammar bias (Li et al., 2019), to achieve maximum performance. For example, Noji et al. (2016) reports a performance difference as high as 0.16 unlabelled attachment score between models with and without grammatical constraints. Following the spirit, our method applies three grammatical constraints and verifies the benefit of the grammatical constraints in LLM-based unsupervised dependency parsing.

4 Method

In this section, we propose a grammatically constrained CMI-based unsupervised parsing method. We apply CMI as the bi-lexical dependence metric and use the sentence distribution derived from a causal LLM as CMI's probability measure. We derive a reliable MH-based CMI estimator incorporating POS information through the POS constraint. We implement a Multi-Try MH (MTMH) sampler (Martino, 2018) to achieve a higher sampling efficiency and a better sample quality. We heuristically incorporate the AC and the FNUH constraint during the MST decoding stage.

4.1 MTMH-based CMI Estimator

First, we introduce an MTMH-based CMI estimator using an LLM-based sentence distribution p. Given a pair of words (x_i, x_j), we compute the CMI value between them as the expected log probability difference between sentences (x_i, x_j, x_{-ij}) where $(x_i, x_j) \sim X_i, X_j|x_{-ij}$ and sentences (x'_i, x'_j, x_{-ij}) where $(x'_i, x'_j) \sim X_i \otimes X_j|x_{-ij}$. Here, $\otimes$ refers to the cartesian product of probability spaces such that $p(X \otimes Y) = p(X)p(Y)$ (i.e., X and Y are independent). We gather samples $\{(x_i^k, x_j^k)\}_k$ from $X_i, X_j|x_{-ij}$ by iteratively performing MTMH sampling for the i-th and the j-th word while keeping the context x_{-ij} unmodified. We then create independent samples $\{(x_i^{k'}, x_j^{k'})\}_k$ using the gathered samples by shuffling among the $\{x_j^k\}_k$ samples while keeping the $\{x_i^k\}_k$ samples unmodified. We estimate the CMI score using the gathered samples and the independent samples, as shown in Equation 3.

$$\begin{aligned}
&I_{ij}^{cmi}(x) \\
&= \mathop{\mathbb{E}}_{(x_i,x_j)\sim X_i,X_j|x_{-ij}} \left[\log \frac{p(x_i|x_j,x_{-ij})p(x_j,x_{-ij})}{p(x_i|x_{-ij})p(x_j,x_{-ij})}\right] \\
&= \mathop{\mathbb{E}}_{\substack{(x_i,x_j)\sim X_i,X_j|x_{-ij} \\ (x'_i,x'_j)\sim X_i\otimes X_j|x_{-ij}}} \left[\log \frac{p(x_i,x_j,x_{-ij})}{p(x'_i|x_{-ij})p(x'_j,x_{-ij})}\right] \\
&= \mathop{\mathbb{E}}_{(x_i,x_j)\sim X_i,X_j|x_{-ij}} [\log p(x_i,x_j,x_{-ij})] \\
&\quad - \mathop{\mathbb{E}}_{(x'_i,x'_j)\sim X_i\otimes X_j|x_{-ij}} [\log p(x'_i,x'_j,x_{-ij})]
\end{aligned} \tag{3}$$

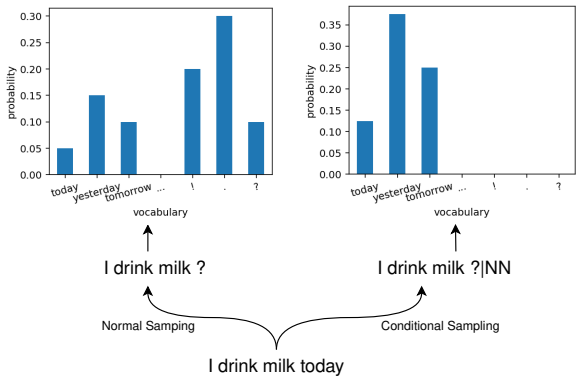

Figure 2: An illustration of the masking process for the conditional sampling.

4.2 Incorporating POS Information through the POS Constraint

We further incorporate POS information into the MTMH sampler using the POS constraint. The POS constraint forces the sampled word to have the same POS as the original word, preventing the MTMH samples' dependency structure from deviating from the original sentence's structure. For example, an unconstrained MTMH sampler might return samples such as "I drink milk !" or "I drink milk ." when sampling for "today" in "I drink milk today". The sampled sentences do not share the same dependency structure as the original sentence. Consequently, identifying dependency structures without the POS constraint becomes more difficult because the CMI score accounts for both the target dependency structure and all structures that can arise throughout the sampling process.

The POS-constrained CMI estimator conditions the CMI on the POS tag (y_i, y_j) of (x_i, x_j). However, no causal LLM provides interfaces to directly condition its sentence distribution on POS tags (i.e., we can not trivially sample from $p(X_i, X_j | x_{-ij}, Y(X_i) = y_i, Y(X_j) = y_j)$). To sample from the conditional distribution, we modify MTMH's proposal distribution such that the probability for illegitimate words is 0 (Figure 2). We refer to the illegitimate word X_i as words violating the POS constraint (i.e., $Y(X_i) \neq y_i$). The modification excludes all illegitimate words from being sampled while keeping the relative probability $\frac{p(x')}{p(x)}$ unchanged between sentences x and x' containing legitimate samples. Because the relative probability is unchanged, we can implicitly renormalize the sentence distribution in the MTMH sampler. As a result, the modification allows for conditioning the sentence distribution on POS tags without explicitly modifying the sentence distribution.

We compute the POS-constrained CMI as Equation 4, given samples $(x_i, x_j) \sim P_j := X_i, X_j | x_{-ij}, Y(X_i) = y_i, Y(X_j) = y_j$ and samples $(x'_i, x'_j) \sim P_m := X_i \otimes X_j | x_{-ij}, Y(X_i) = y_i, Y(X_j) = y_j$. We show in Appendix A.1 that the computation would yield the conditional CMI score, assuming that POS is unambiguous given the entire sentence. The assumption holds for the vast majority of natural language sentences.

$$I^{cmi}_{ij}(x|y_i, y_j) = \mathop{\mathbb{E}}_{(x_i,x_j)\sim P_j} [\log p(x_i, x_j, x_{-ij})] - \mathop{\mathbb{E}}_{(x'_i,x'_j)\sim P_m} [\log p(x'_i, x'_j, x_{-ij})] \quad (4)$$

4.3 Decoding Dependencies with the AC and the FNUH Constraint

We heuristically apply two grammatical constraints (the Adjacent-Connect (AC) constraint and the Function Word UnHeading (FNUH) constraint) and apply Prim's algorithm during the decoding stage. The AC constraint injects a locality bias by forcing a word to connect with its right neighbor when the word has a low CMI score with the rest of the sentence. The constraint comes into effect when the cumulative CMI of the word is smaller than a preset threshold (i.e., $\sum_j I^{cmi}_{ij}(x) < \tau$). The FNUH constraint injects a structural bias by preventing function words from being a syntactic head in the predicted dependency tree. This constraint exploits the structural bias that Universal Dependencies rarely use function words as a head and was applied in previous research (Noji et al., 2016). In undirected dependency parsing, being a headword means the word has a degree larger than one in the decoded dependency structure. We enforce the constraint by gradually discounting the CMI score related to the function word that violates the constraint. The FNUH constraint can only suppress high CMI scores assigned in wrong dependency structures (type I error) but is unable to uncover dependencies that are not detected by CMI (type II error).

5 Experiments and Results

5.1 Experiment Setup

We conduct experiments using the Universal Dependency (UD) (Nivre et al., 2020) datasets in

five languages (i.e., English (EN), German (DE), French (FR), Spanish (ES), and Russian (RU)) for analysis. We use the EWT (Bies, Ann et al., 2012) section for English and the GSD section (Nivre et al., 2020) for the other four languages. We carry out the evaluation over 10-word subsets of the respective datasets, which contain sentences with at most ten words without punctuations. This setting is common in previous unsupervised parsing research (Klein and Manning, 2004; Cohen and Smith, 2009). We additionally include the WSJ10 dataset and the full English Parallel Universal Dependency (PUD) dataset for comparison with state-of-the-art methods, as done in Wu et al. (2020).

We apply the unbiased Cohen's d metric (Hedges, 1981), an effect size metric, to measure the correlation strength between dependence scores and syntactic dependencies. A higher d value indicates a stronger correlation (Gibson, 2015). We adopt the Unlabelled Undirected Attachment Score (UUAS) (Hewitt and Manning, 2019) for evaluating the parsing performance. UUAS is our primary evaluation metric because the CMI score is symmetric in definition. Consequently, the CMI model cannot recover directed dependencies. We measure only for dependencies connecting actual words (i.e., we exclude the root dependency and any dependencies connecting to punctuations). This setting aligns with the evaluation principle in previous works (Klein and Manning, 2004; Noji et al., 2016) where all punctuations are removed.

For the CMI estimation, we use the multilingual BERT model (Devlin et al., 2018) for the MTMH sampler's proposal distribution and the multilingual GPT model (Shliazhko et al., 2022) for the sampler's target distribution. We use the Perturbed-Masking method (Wu et al., 2020) (PM) as the baseline method [1]. We apply the same multilingual BERT model to the CMI and PM methods. We also include results for PM(bbu), a PM variant using the monolingual `bert-base-uncased` model (Devlin et al., 2018), to align with previous evaluation settings. We refer to the CMI variant using the POS constraint as CMI, as the constraint is an integral part of the CMI score estimation process.

5.2 CMI-Syntactic Dependency Correlation

Figure 3 compares the CMI score between syntactic dependencies and non-dependencies for dependencies with lengths up to 6. The dependency length is the number of words between the two words connected by the dependency. We observe a positive correlation: the CMI score tends to be higher for the syntactic dependency than for the non-dependency. This result highlights the CMI score's ability to separate the two types of dependencies. However, the CMI score's magnitude decreases monotonically with the dependency length such that the CMI score for the long syntactic dependency is, on average, lower than the CMI score for the short non-dependency. This result indicates a weaker statistical correlation for words connected by the long syntactic dependency.

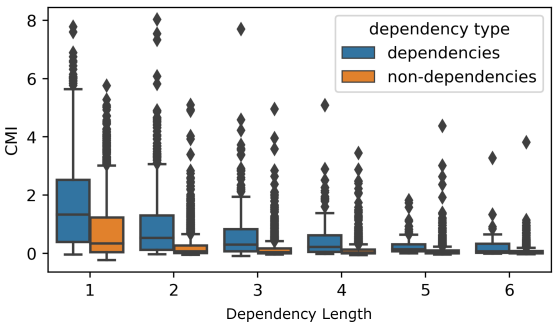

Figure 3: CMI scores for syntactic dependencies and non-dependencies binned by dependency lengths. The score is estimated over the 10-word English EWT dataset.

	EN	DE	FR	RU	ES
	d				
CMI	1.03	1.31	1.26	1.24	1.24
PM	1.05	1.13	1.25	1.18	1.24
PM(bbu)	1.17	-	-	-	-
	$\bar{d}$				
CMI	0.81	1.21	0.8	0.82	1.03
PM	0.74	0.97	0.73	0.55	0.64
PM(bbu)	0.81	-	-	-	-

Table 1: d and $\bar{d}$ metrics computed for the CMI and PM scores. - indicates that the model can not run on the corresponding dataset.

5.3 Effect Size

Table 1 compares the correlation strength of the CMI score with the strength of the PM score using two effect size metrics (d and $\bar{d}$). The d metric computes the Cohen's d value for all dependencies as a single group, whereas the $\bar{d}$ metric computes the average Cohen's d value for equal-length dependencies. The table suggests that the CMI score generally exhibits a stronger correlation across languages than the PM score, as evidenced by the

[1] We removed a buggy softmax implementation in decoding because we found that parsing with the raw PM score yields a better performance.

Models	EN	DE	FR	RU	ES	Mean
AC	0.497	0.480	0.513	0.579	0.530	0.519
CMI w/o GR	0.591	0.605	0.595	0.603	0.594	0.597
CMI w/o GR POS	0.556	0.563	0.565	0.575	0.569	0.565
PM w/o GR	0.540	0.563	0.548	0.576	0.603	0.566
PM(bbu)w/o GR	0.576	-	-	-	-	-

Table 2: UUAS of the CMI model and the PM model in 5 languages. GR and POS indicate the use of gold-root and Part-Of-Speech information, respectively. - indicates that the model can not run on the respective dataset. The mean column shows the average UUAS across all languages.

significantly higher $\bar{d}$ values. Both the CMI and the PM score appear to have a weaker correlation than the monolingual PM(bbu) score in English. As we will see in the next section, the CMI and the PM(bbu) models have higher parsing performance than the PM model, as the effect size metrics predicted. However, the CMI model outperforms the PM(bbu) model in parsing performance despite the seemingly stronger correlation of the PM(bbu) score.

5.4 Parsing Performance Against Baselines

5.4.1 UUAS Comparison

Table 2 compares the CMI parsing model without grammatical constraints applied in the decoding stage and the PM model without the gold-root (GR) information injection. We include an Adjacent-Connect (AC) baseline, which forms a dependency tree by connecting adjacent words. The AC baseline is the undirected variant of the trivial right-branching baseline (Klein and Manning, 2004) that performs strongly in unsupervised dependency parsing. We convert the directed dependency output from the PM model to undirected ones when evaluating the UUAS for the PM model.

As shown in the table, the CMI model outperforms the baseline PM model and the CMI w/o POS model by an average of 0.031 UUAS and 0.032 UUAS, respectively. In Spanish, where the CMI model underperforms the PM model, the performance gap is less significant than in other languages, where the CMI model outperforms the PM model. The CMI model also outperforms the monolingual PM(bbu) model by 0.015 UUAS in the English dataset. This result demonstrates the effectiveness of the POS constraint and establishes the CMI model as a strong LLM-based unsupervised dependency parser.

Relations	EN	DE	FR	RU	ES	Mean
nsubj	0.092	0.004	0.130	0.162	0.110	0.100
obj	0.063	0.143	0.119	0.180	-0.060	0.089
iobj	-	-0.077	-	0.091	-0.234	-0.073
ccomp	0.162	-0.083	-	-	-	0.039
xcomp	0.278	0.067	-0.143	0.048	-0.188	0.012

Table 3: UUAS difference between the CMI w/o GR model and the PM w/o GR model for core relations. The table only shows relations with more than ten occurrences in the dataset. `csubj` is excluded because none of the datasets contains more than ten occurrences.

5.4.2 Comparing Core Dependency Extractions

Table 3 shows the UUAS difference between the CMI w/o GR model and the PM w/o GR model on five core dependency relations. The CMI model performs strongly in retrieving core dependencies, outperforming the PM model by 0.033 mean UUAS on average across the five relations. The CMI model performs exceptionally well on dependencies of the `nsubj` and `obj` relations. The two core relations typically connect words with richer lexical interactions than the other three core relations. We denote a dependency relation as lexical-interaction-rich if the words connected by the dependency tend to have a strong correlation pattern and a relation as lexical-interaction-sparse if the words have a weak correlation pattern. For example, relations such as `obj` and `compound` would have richer lexical interactions than relations such as `cop`. Beyond the core relations, we found the CMI model performing strongly on lexical-interaction-rich relations such as `nmod` (avg. 0.09 difference) and `compound` (avg. 0.12 difference). This analysis indicates the CMI model's strength in retrieving lexical-interaction-rich relations.

5.5 CMI's Problem in Parsing

Figure 4 illustrates at least two problems of the CMI model. Firstly, the CMI model performs weakly in capturing dependencies that involve sparse lexical interactions. This weakness is ev-

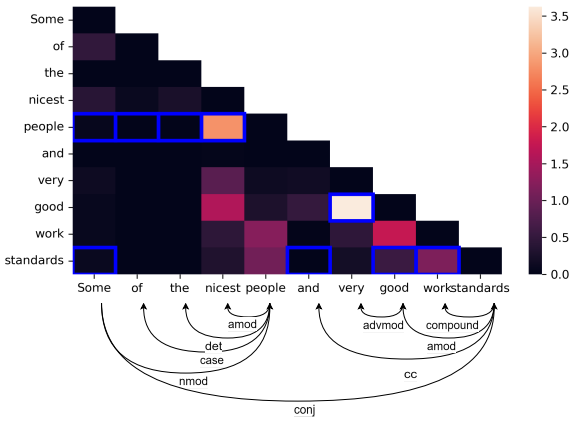

Figure 4: A negative example where the CMI model fails to capture many syntactic dependencies.

	EN	DE	FR	RU	ES
Lexical-Interaction-Sparse Relations					
cop	0.322	0.348	0.333	-	0.515
case	0.556	0.533	0.547	0.696	0.435
mark	0.363	0.370	0.250	-	0.500
Lexical-Interaction-Rich Relations					
obj	0.695	0.619	0.644	0.620	0.640
nsubj	0.538	0.565	0.580	0.619	0.524
nmod	0.500	0.566	0.620	0.581	0.628
Model UUAS	0.591	0.605	0.595	0.603	0.594

Table 4: UUAS of the CMI model for lexical-interaction-rich relations (`nmod`, `obj`, and `nsubj`) and lexical-interaction-sparse relations (`cop`, `case`, and `mark`). Only UUAS for relations with more than ten occurrences are shown. The "Model UUAS" entry shows the UUAS for all dependencies.

idenced by the low CMI score for dependencies with sparse lexical interactions. Table 4 compares the UUAS of the CMI model between the lexical-interaction-rich and the lexical-interaction-sparse relations. The CMI model tends to produce higher than overall UUAS for the lexical-interaction-rich relations (`obj`, `nsubj`, and `nmod`) and produce lower UUAS for the lexical-interaction-sparse relations (`cop`, `case`, and `mark`). Other lexical-interaction rich relations, such as `amod`, can reach a mean UUAS up to 0.802 across the five languages. Secondly, the CMI model reacts to words with similar meanings (e.g., the pair "nicest" and "good" in Figure 4 gets a high CMI), yet the words are syntactically unrelated. The two problems indicate a limitation in CMI-based methods in solving the unsupervised dependency parsing problem.

5.6 Effect of the AC and the FNUH Constraint

Table 5 compares the UUAS of the CMI model with different levels of grammatical constraints

Models	EN	DE	FR	RU	ES
CMI w/ AC FNUH	0.645	0.662	0.668	0.644	0.679
CMI w/ AC	0.615	0.62	0.625	0.637	0.627
CMI	0.591	0.605	0.595	0.603	0.594

Table 5: UUAS of the CMI model with different levels of grammatical constraints.

	EN	DE	FR	RU	ES
Lexical-Interaction-Sparse Relations					
cop	-0.181	-0.030	0.056	-	0.000
case	-0.134	0.033	-0.035	-0.070	-0.031
mark	-0.069	0.000	0.000	-	0.000
Lexical-Interaction-Rich Relations					
obj	0.049	0.076	0.085	0.020	0.100
nsubj	0.143	0.117	0.099	0.000	0.061
nmod	0.277	0.133	0.167	0.035	0.198

Table 6: UUAS difference between the CMI w/ FNUH AC model and CMI w/ AC model on three lexical-interaction-sparse relations and three lexical-interaction-rich relations. The table shows only relations with ≥ 10 occurrences.

applied at the decoding stage. Both the AC and the FNUH constraint contribute positively towards the parsing performance. The two constraints, together, result in 0.062 UUAS improvement on average. The FNUH is the more influential constraint, resulting in 0.034 UUAS improvements on average. Further analysis (Table 6) shows FNUH's benefit in recovering relations with rich lexical interactions. The `obj`, `nsubj`, and `nmod` relations obtain on average 0.104 and up to 0.277 UUAS improvement by applying the FNUH constraint. This result indicates that the FNUH can effectively reveal lexical-interaction-rich dependencies detected by CMI. Nonetheless, there are drawbacks in applying the FNUH constraint, namely, UUAS loss in relations with sparse lexical interactions (`cop`, `case`, and `mark`). The drawback is expected because those relations connect to a function word and can be a victim of the FNUH constraint. In conclusion, the two grammatical constraints are effective in recovering more syntactic dependencies and are generally applicable to many languages.

5.7 Parsing Performance Against SOTA

Table 7 compares the parsing performance of the CMI model with the AC and FNUH constraint with the performance of three LLM-based and three grammar-based models. The comparison is carried out over the test set of the respective datasets. We apply the gold-root information to the PM model and apply the FNUH constraint to the DMV and LCDMV model, following Wu et al. (2020) and

	EN	DE	FR	RU	ES	EN-PUD	WSJ10(S)	WSJ10(C)	Mean
Multilingual LLM-Based Methods									
CMI w/ AC FNUH	0.634	0.563	0.681	0.637	0.726	0.541	0.589	0.548	0.615
PM (Wu et al., 2020)	0.568	0.582	0.582	0.591	0.606	0.506	0.592	0.462	0.575
MB (Zhang and Hashimoto, 2021)	0.521	0.480	0.544	0.536	0.570	0.473	0.552	0.576	0.525
Grammar-Based Methods									
DMV (Klein and Manning, 2004)	0.612	0.604	0.592	0.678	0.705	0.484	0.597	0.555	0.603
LCDMV (Noji et al., 2016)	0.658	0.626	0.642	0.731	0.608	0.554	0.614	0.578	0.626
Joint (Yang et al., 2020)	0.650	0.564	0.738	0.420	0.662	0.556	0.704	0.792	0.636
Monolingual LLM-based Methods									
PM(bbu)	0.602	-	-	-	-	0.511	0.603	0.53	-
MB(bbc)	0.352	-	-	-	-	0.495	0.586	0.561	-
SSUD (Jian and Reddy, 2023)*	-	-	-	-	-	0.464	0.576	-	-

Table 7: Model UUAS on the 10-word test sets, the English PUD dataset, and the WSJ10 dataset. - indicates that the model cannot run on the corresponding dataset. WSJ10(S) and WSJ10(C) refer to the WSJ10 corpus annotated in the Stanford dependency (de Marneffe et al., 2006) and the Collins dependency (Collins, 2003) format, respectively. * indicates results from the original paper. MB(bbc) refers to the MB model using the `bert-base-cased` model. The mean column shows the average UUAS across the eight datasets.

Noji et al. (2016) respectively. We initialize the Joint model using the LCDMV model's prediction on the universal dependency treebanks and the HDP-DEP (Naseem et al., 2010) model's prediction on the WSJ10 datasets.

The table shows that the CMI model outperforms the PM and the MB models by 0.04 UUAS and 0.09 UUAS across the eight datasets. The CMI model outperforms the monolingual PM and MB models in three out of four datasets. Compared to grammar-based models, the CMI model outperforms the similarly constrained DMV model but underperforms the LCDMV and the Joint model. However, the LCDMV and the Joint model are more strongly constrained than the CMI model. The LCDMV model applies a maximum depth constraint to center-embedding structures in addition to the FNUH constraint. The Joint model inherits the depth constraint from the LCDMV model or dependency rule constraints from the HDP-DEP model during its parameter initialization process. This result, on the one hand, establishes the CMI model as a strong LLM-based unsupervised dependency parsing model. On the other hand, the result underscores the significance of grammatical constraints in LLM-based unsupervised parsing, as supported by prior research on grammar-based unsupervised dependency parsing.

6 Discussion

In Section 5, we saw grammar-based models outperforming the CMI model after applying all three grammatical constraints. Nonetheless, we believe the grammar-based models are not superior substitutes for the CMI-based models. The two models are on opposite ends of a spectrum. Our experiments have pointed out the strong performance of the CMI model in retrieving dependency relations with rich lexical interactions, which is in line with Yuret (1998)'s finding. Although grammar-based models can benefit from lexical information (Han et al., 2017), they struggle to utilize the rich lexical interaction encoded in LLMs (Han et al., 2020). The CMI-based models and the grammar-based models are complementary. We believe that combining the strength of the CMI and grammar-based models can lead to a more robust unsupervised parsing method.

7 Conclusion

In this paper, we applied CMI, an interpretable bi-lexical dependence metric, to unsupervised dependency parsing and proposed a reliable MTMH-based CMI estimator. We incorporated POS information into the CMI estimation process through the POS constraint and further integrated the AC and FNUH constraints at the decoding stage. The correlation analysis suggests that the CMI score positively correlates with the syntactic dependency and has a stronger correlation with the syntactic dependency than baseline scores. The comparison with the baseline models and the ablation analysis confirmed the effectiveness and applicability of the three grammatical constraints in LLM-based unsupervised dependency parsing across five languages. Analysis by dependency relations indicates that the CMI model performs strongly on relations involving rich lexical interactions but performs poorly on relations involving sparse lexical interactions. The weakness in retrieving lexical-interaction-sparse relations suggests a limitation in CMI-based unsupervised parsing methods. The comparison with

state-of-the-art models establishes the CMI model as a strong LLM-based model, which outperforms LLM-based and similarly constrained grammar-based models but underperforms the more strongly constrained model.

8 Limitations

One issue with the CMI method is the amount of computation needed to estimate the CMI score. For example, we need 40 GPU hours on A100 GPU to evaluate all CMI scores for a 10-word UD dataset. The computation complexity arises from two sources: the MTMH sampling process and $O(n^2)$ complexity for computing the CMI score for all word pairs. The MTMH algorithm iteratively performs sampling over x_i and x_j. At step k, the algorithm needs to evaluate the proposal distribution $q(X_i|x_{-i}^{k-1})$ and compute the sentence probability for $p(x_i^{k'}, x_{-i}^{k-1})$. Although the current implementation can evaluate the above two in $O(1)$ time, each evaluation is expensive due to the large number of parameters in causal LLMs. This complexity is further compounded by the need for many iterations between samples to minimize autocorrelation.

9 Acknowledgement

This research was supported by Grant-in-Aid for JSPS Fellows, Numbered 23KJ0565.

References

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

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. Variational inference for grammar induction with prior knowledge. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 1–4, Suntec, Singapore. Association for Computational Linguistics.

Michael Collins. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637.

Thomas M. Cover and Joy A. Thomas. 1991. Elements of information theory.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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.

Douglas Gibson. 2015. Effect size as the essential statistic in developing methods for mtbi diagnosis. *Frontiers in Neurology*, 6.

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.

Wenjuan Han, Yong Jiang, and Kewei Tu. 2017. Dependency grammar induction with neural lexicalization and big training data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1683–1688, Copenhagen, Denmark. Association for Computational Linguistics.

W. K. Hastings. 1970. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Larry V Hedges. 1981. Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6:107 – 128.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for 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.

Jasper Jian and Siva Reddy. 2023. Syntactic substitutability as unsupervised dependency syntax. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2341–2360, Singapore. 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.

Bowen Li, Jianpeng Cheng, Yang Liu, and Frank Keller. 2019. Dependency grammar induction with a neural variational transition-based parser. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6658–6665. AAAI Press.

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.

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

Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 2019. CGMH: constrained sentence generation by metropolis-hastings sampling. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6834–6842. AAAI Press.

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.

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.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *CoRR*, abs/2204.07580.

Sameer Singh, Michael Wick, and Andrew McCallum. 2012. Monte Carlo MCMC: Efficient inference by approximate sampling. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1104–1113, Jeju Island, Korea. Association for Computational Linguistics.

Noah A. Smith and Jason Eisner. 2006. Annealing structural bias in multilingual weighted grammar induction. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 569–576, Sydney, Australia. Association for Computational Linguistics.

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.

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.

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.

Deniz Yuret. 1998. *Discovery of linguistic relations using lexical attraction*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

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 Conditional CMI computation

Proposition 1. *$I_{ij}^{cmi}(x|y_i, y_j)$ can be computed as Equation 4 assuming the unambiguity of POS given the full sentence.*

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

$$I_{ij}^{cmi}(x|x_{-ij}, y_i, y_j)$$

$$:= \mathbb{E}_{X_iX_j|c}\left[\log \frac{p(x_i, x_j, x_{-ij}|y_i, y_j)}{p(x_i|x_{-ij}, y_i, y_j)p(x_j, x_{-ij}|y_i, y_j)}\right] \quad (5)$$

$$= \mathbb{E}_{\substack{(x_i,x_j)\sim X_iX_j|c \\ (x_i',x_j')\sim X_i\otimes X_j|c}}\left[\log \frac{p(x_i, x_j, x_{-ij}|y_i, y_j)}{p(x_i', x_j', x_{-ij}|y_i, y_j)}\right] \quad (6)$$

$$= \mathbb{E}_{\substack{(x_i,x_j)\sim X_iX_j|c \\ (x_i',x_j')\sim X_i\otimes X_j|c}}\left[\log \frac{p(x_i, x_j, x_{-ij})\mathbb{1}_{Y(x_i,x_j)=y_i,y_j}}{p(x_i', x_j', x_{-ij})\mathbb{1}_{Y(x_i',x_j')=y_i,y_j}}\right] \quad (7)$$

$$= \mathbb{E}_{\substack{(x_i,x_j)\sim X_iX_j|c \\ (x_i',x_j')\sim X_i\otimes X_j|c}}\left[\log \frac{p(x_i, x_j, x_{-ij})}{p(x_i', x_j', x_{-ij})}\right] \quad (8)$$

□

We apply the unambiguous POS assumption in Equation 6, converting the conditional probability to the product of an indicator function and the sentence probability. Because the indicator function is satisfied by the condition c, the indicator function will always be 1.

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

They proposed a formulation of "conditional mutual information" (Equation 9)

$$I_{ij}^{ZH}(x) = \mathop{\mathbb{E}}_{X_iX_j|x_{-ij}}\Big[\log p(x_i|x_j, x_{-ij}) - \log \mathop{\mathbb{E}}_{X_j|x_i,x_{-ij}} p(x_i|x_j, x_{-ij})\Big] \quad (9)$$

We prove the following propositions

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

Proof.

$$(9) = \mathop{\mathbb{E}}_{X_i|x_{-ij}}\Big[\mathop{\mathbb{E}}_{X_j|x_i,x_{-ij}} \log p(x_i|x_j, x_{-ij}) - \log \mathop{\mathbb{E}}_{X_j|x_i,x_{-ij}} p(x_i|x_j, x_{-ij})\Big] \quad (10)$$

$$\leq \mathop{\mathbb{E}}_{X_i|x_{-ij}}\Big[\mathop{\mathbb{E}}_{X_j|x_i,x_{-ij}}\Big[\log p(x_i|x_j, x_{-ij}) - \mathop{\mathbb{E}}_{X_j|x_i,x_{-ij}} \log p(x_i|x_j, x_{-ij})\Big]\Big] \quad (11)$$

$$= 0 \quad (12)$$

□

Proposition 3. *Two statistically independent random variables 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

$$I^{ZH}(X_i; X_j) = (2 * 0.5)\Big[(0.5 * 2)\log 0.5 \quad (13)$$

$$- \log(0.5 * 2 * 0.5)\Big] \quad (14)$$

$$= 0 \quad (15)$$

□