Hidden Schema Networks

R. J. Sánchez[†‡] (sanchez@bit.uni-bonn.de), L. Conrads[†‡], P. Welke[†*], K. Cvejoski[†≀], C. Ojeda[§]

[†] Lamarr-Institute for Machine Learning and Artificial Intelligence [‡] University of Bonn [≀]Fraunhofer-Institute for Intelligent Analysis and Information Systems (IAIS) [*] TU Wien [§] University of Potsdam

LAMARR
Institute for Machine Learning
and Artificial Intelligence

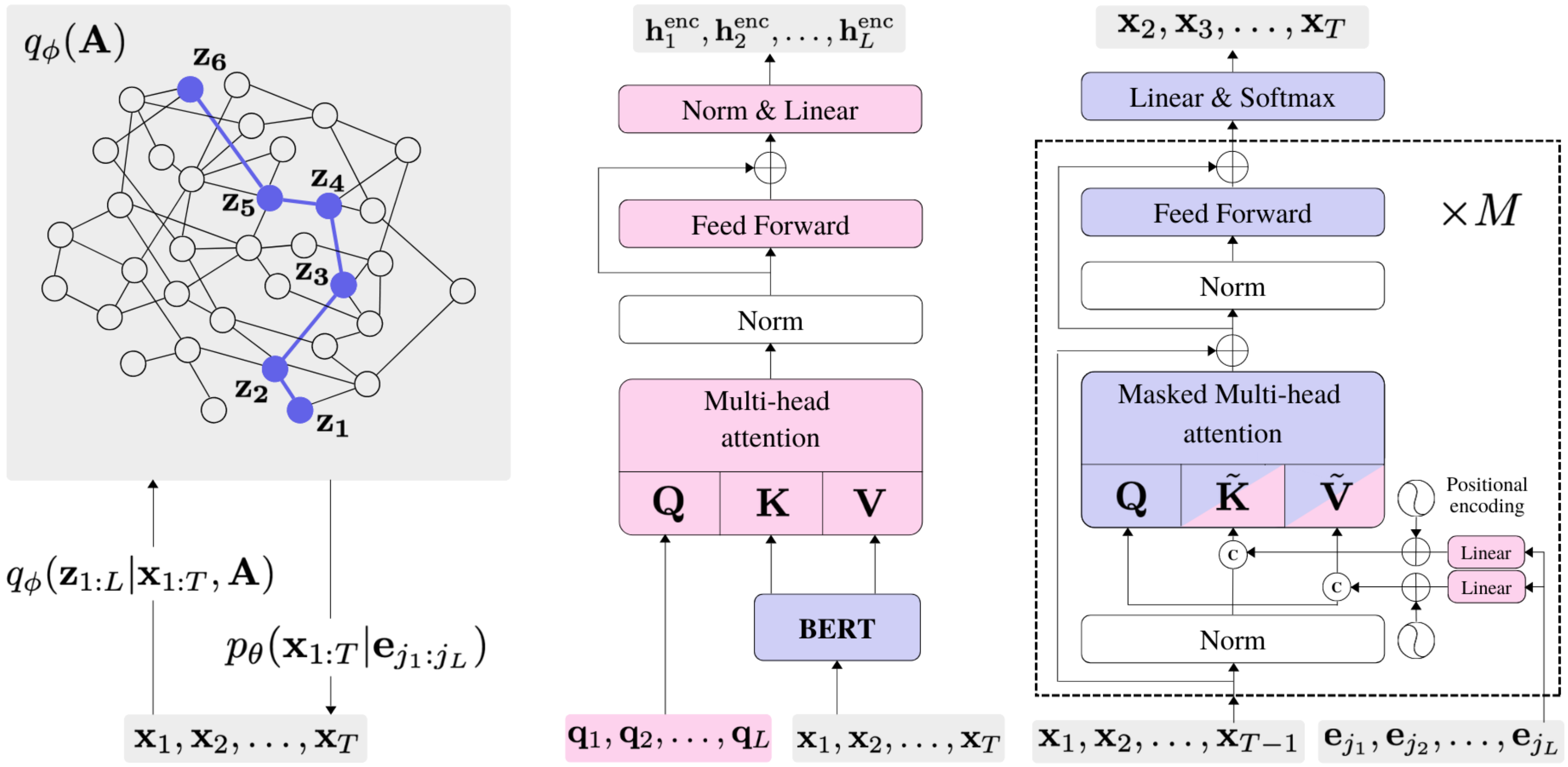
Abstract

Large, pretrained language models infer powerful representations that encode rich semantic and syntactic content, albeit implicitly. In this work we introduce a novel neural language model that enforces, via inductive biases, explicit relational structures which allow for compositionality onto the output representations of pretrained language models. Specifically, the model encodes sentences into sequences of symbols (composed representations), which correspond to the nodes visited by biased random walkers on a global latent graph, and infers the posterior distribution of the latter. We first demonstrate that the model is able to uncover ground-truth graphs from artificially generated datasets of random token sequences. Next, we leverage pretrained BERT and GPT-2 language models as encoder and decoder, respectively, to infer networks of symbols (schemata) from natural language datasets. Our experiments show that (i) the inferred symbols can be interpreted as encoding different aspects of language, as e.g. topics or sentiments, and that (ii) GPT-like models can effectively be conditioned on symbolic representations. Finally, we explore training autoregressive, random walk "reasoning" models on schema networks inferred from commonsense knowledge databases, and using the sampled paths to enhance the performance of pretrained language models on commonsense *If-Then* reasoning tasks.

Motivation and Goals

- Large language models (LLM) struggle to solve tasks that require non-language-specific skills, like formal and commonsense reasoning.
- We seek to translate the linguistic knowledge encoded by LLM into agnostic and unsupervised *representations for reasoning*.
- Contrary to Change-of-Thought, we define reasoning in representation space *à la* Fodor, as a type of *ordered mental expressions*.

The Hidden Schema Model (HSN)


$q_\phi(\mathbf{A})$
$\mathbf{z_6}$
$\mathbf{z_5}$
$\mathbf{z_4}$
$\mathbf{z_3}$
$\mathbf{z_2}$
$\mathbf{z_1}$
$q_\phi(\mathbf{z}_{1:L}|\mathbf{x}_{1:T}, \mathbf{A})$
$p_\theta(\mathbf{x}_{1:T}|\mathbf{e}_{j_1:j_L})$
$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$
$\mathbf{h}_1^{\text{enc}}, \mathbf{h}_2^{\text{enc}}, \dots, \mathbf{h}_L^{\text{enc}}$
Norm & Linear
Feed Forward
Norm
Multi-head attention
Q
K
V
BERT
$\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_L$
$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$
$\mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_T$
Linear & Softmax
Feed Forward
$\times M$
Norm
Masked Multi-head attention
Q
$\tilde{\text{K}}$
$\tilde{\text{V}}$
Positional encoding
c
c
Linear
Linear
Norm
$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{T-1}$
$\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, \dots, \mathbf{e}_{j_L}$

Generative Model

$$p_\theta(\mathbf{x}_{1:T}, \mathbf{A}) = \sum_{\mathbf{z}_{1:L}} p_\theta(\mathbf{x}_{1:T}|\mathbf{e}_{j_1:j_L}) p(\mathbf{z}_{1:L}|\mathbf{A}) p(\mathbf{A}). \tag{1}$$

Posterior over random walks (encoder model)

$$q_\phi(\mathbf{z}_{1:L}|\mathbf{x}_{1:T}, \mathbf{A}) = \left(\prod_{i=1}^{K} \rho_i(\mathbf{x}_{1:T}, \phi)^{z_1^i}\right) \cdot \prod_{i=2}^{L} \left(\prod_{j=1}^{K}\prod_{k=1}^{K} \left(Q_{k,j}^{[i-1]}(\mathbf{x}_{1:T}, \mathbf{A}, \phi)\right)^{z_i^k z_{i-1}^j}\right), \tag{2}$$

where $\boldsymbol{\rho}(\mathbf{x}_{1:T}, \phi) = \text{softmax}(\mathbf{h}_1^{\text{enc}})$ and

$$Q_{k,j}^{[i]}(\mathbf{x}_{1:T}, \mathbf{A}, \phi) = \frac{f_k^{[i]}(\mathbf{x}_{1:T}, \phi)\, A_{kj}}{\sum_m f_m^{[i]}(\mathbf{x}_{1:T}, \phi)\, A_{mj}}, \text{with } \mathbf{f}^{[1]}, \dots, \mathbf{f}^{[L-1]} = \exp(\mathbf{h}_{2:L}^{\text{enc}}).$$

Posterior over global graphs

$$q_\phi(\mathbf{A}) = \prod_{i,j} p_\phi(\mathbf{e}_i, \mathbf{e}_j)^{a_{ij}} \left(1 - p_\phi(\mathbf{e}_i, \mathbf{e}_j)\right)^{1-a_{ij}} \text{ where } p_\phi(\mathbf{e}_i, \mathbf{e}_j) = \text{sigmoid}(g_\phi(\mathbf{e}_i, \mathbf{e}_j)), \tag{3}$$

Inferring Ground Truth Networks

- Given a ground-truth graph $\mathcal{G}^*$, we assign one random bag of tokens to each node of $\mathcal{G}^*$.
- We sample a set of uniform random walks on $\mathcal{G}^*$, and sample a random token from each node along the walks.
- The task is to infer $\mathcal{G}^*$ from the random token sequences.

Graph $\mathcal{G}^*$	ROC AUC	$\lvert\lvert\mathcal{G}^* - \mathcal{G}\rvert\rvert_F$	$\lvert\lvert\mathcal{G}^{\text{rand}} - \mathcal{G}\rvert\rvert_F$	N. edges($\mathcal{G}$)	N. edges($\mathcal{G}^*$)
Barabasi	0.989 ± 0.001	17 ± 2	26 ± 1	1360 ± 104	291
Erdos	0.94 ± 0.06	36.8 ± 0.8	44 ± 2	3131 ± 156	2092

Results on ground-truth random graphs inference. $\mathcal{G}^*$ ($\mathcal{G}$) labels ground-truth (discovered) graph. $||\cdot||_F$ labels Frobenius norm. Error bars are computed from 10 random model initializations.

Links

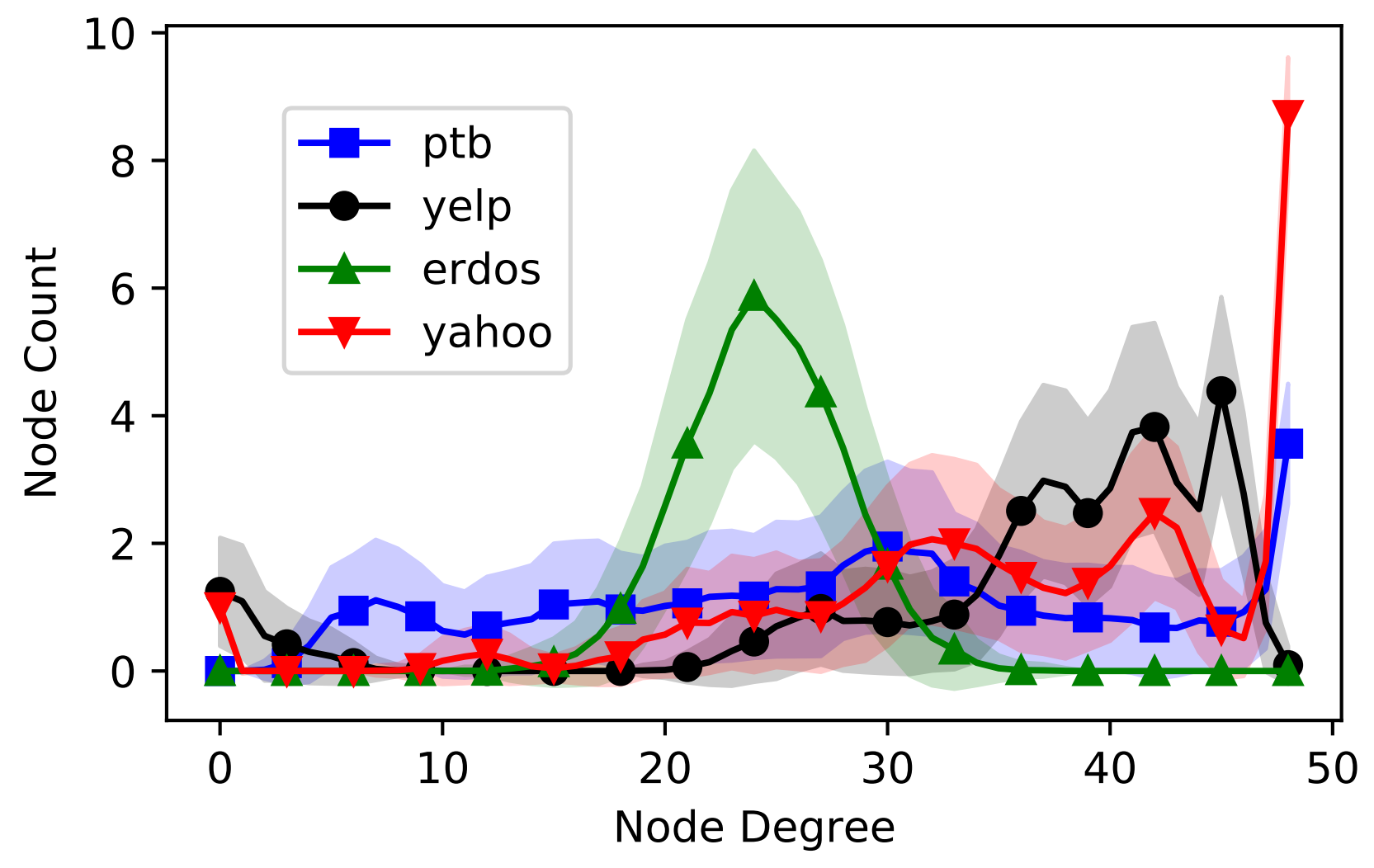
(a) Paper

(b) Source code

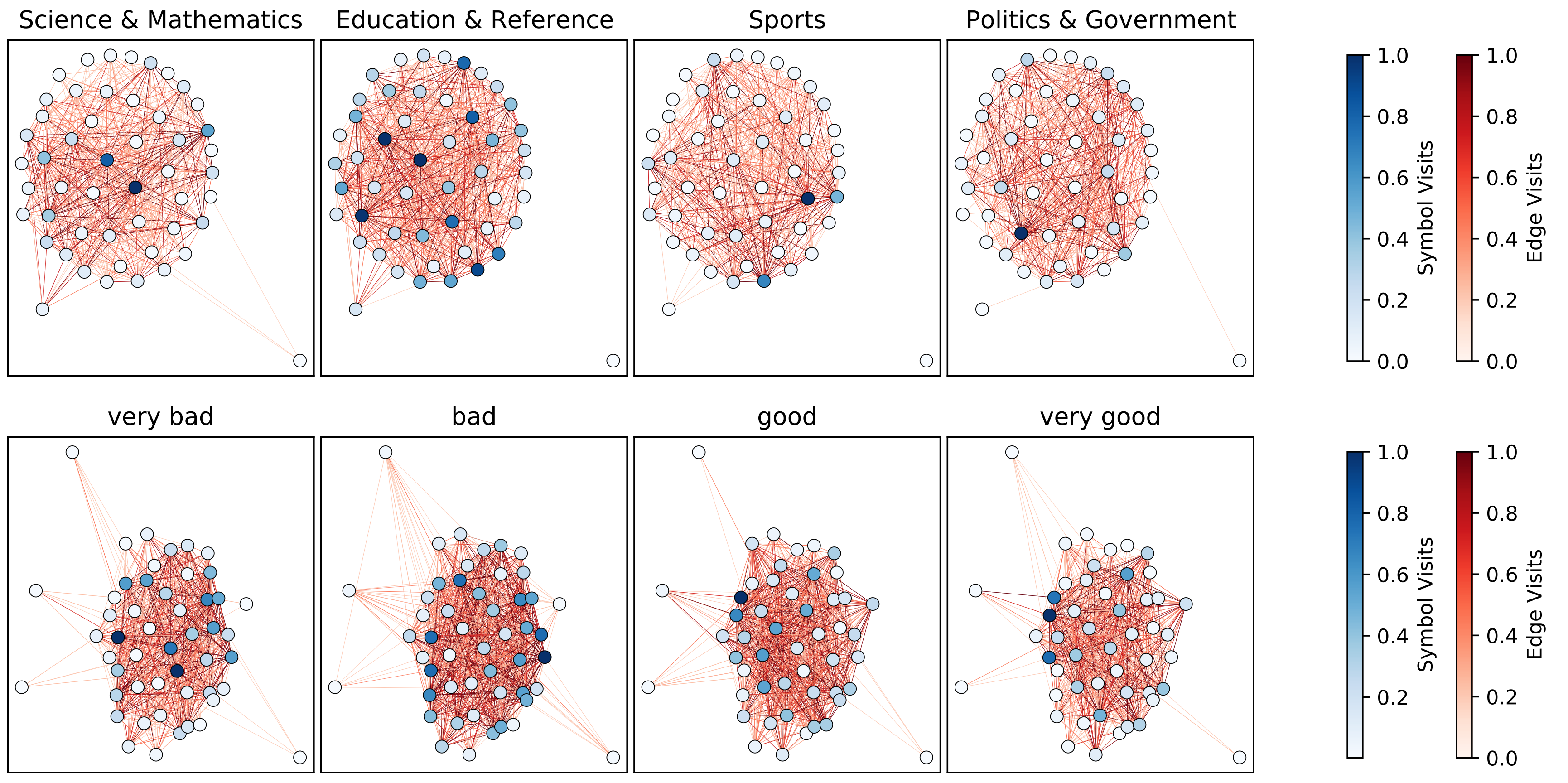
Inferring HSN from Natural Language

	PTB		YAHOO		YELP	
Model	PPL	MI	PPL	MI	PPL	MI
GPT2	24.23	–	22.00	–	23.40	–
iVAE$_{\text{MI}}$	53.44	12.50	47.93	10.70	36.88	11.00
Optimus$_{\text{A}}$	23.58	3.78	22.34	5.34	21.99	2.54
Optimus$_{\text{B}}$	35.53	8.18	29.92	9.18	24.59	9.13
HSN$_{(50,\ 5)}$	22.47	9.50	**20.99**	10.42	**19.72**	10.04
HSN$_{(50,\ 20)}$	30.38	**26.06**	22.84	**22.81**	21.60	**24.93**
HSN$_{(100,\ 5)}$	**20.25**	9.30	21.01	11.21	19.82	10.20
HSN$_{(100,20)}$	25.48	23.77	21.98	16.13	21.13	18.85

Perplexity (PPL) (lower is better) and mutual information (MI). GPT2 and Optimus$_{\text{A, B}}$ results were extracted from Li. et al (EMNLP 2020). Optimus$_{\text{A, B}}$ label models with best PPL and MI, respectively (with $\lambda = 0.05, 1$). iVAE$_{\text{MI}}$ was taken from Fang et al. (EMNLP 2019). We sampled 100 (10) random walks (graphs) to estimate the PPL. End-of-sequence tokens are kept during evaluation.

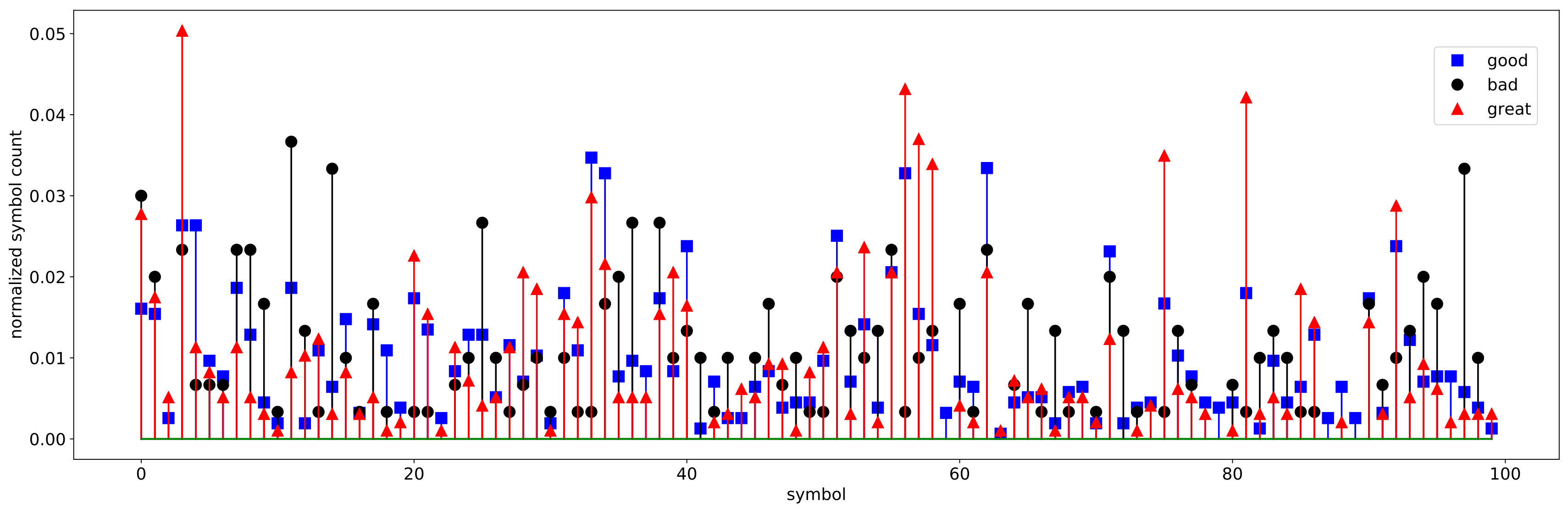


Empirical degree distributions of inferred schema networks against that of an Erdös-Rényi graph with $p = 0.5$. Results correspond to HSN$_{(50,\ 5)}$. The graphs are sampled 500 times. Note that HSN differ from simple random graphs.



Schema distributions inferred from HSN$_{(50,\ 5)}$ for four subsets of the Yahoo (top) and Yelp (bottom) corpora. The node positions in the figure are consistent among labels and were computed using a force-directed embedding of the global graph $\mathcal{G}$.

Which Symbols do Words Attend to? A Preliminary Study on Yelp Reviews



Distribution of most attended symbols when generating tokens *good*, *bad*, *great* for HSN(100, 5) trained on the Yelp data set. The decoder attention matrices between symbols and output are averaged over all attention heads for layer 1 of the decoder network. **Kullback-Leibler divergences**: KL(good, bad) = 0.807, KL(good, great) = 0.336 and KL(great, bad) = 1.227.

Commonsense Reasoning

Subject	Relation	Object
PersonX makes PersonY's coffee	xIntent	PersonX wanted to be helpful

TASK**:** given **Subject + Relation** generate **Object**

Reasoning with **Hidden Schemata**

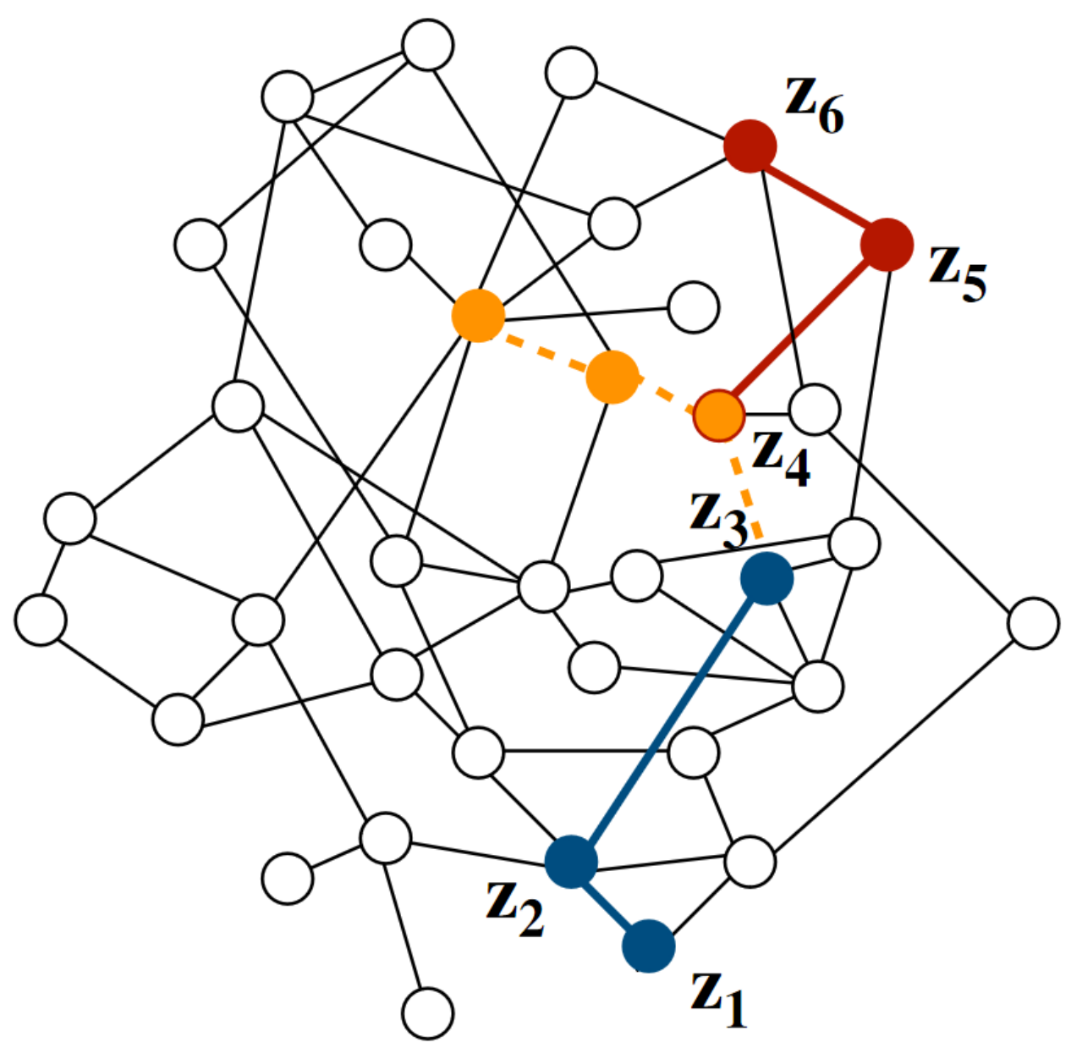
1. Encode **s + r + o** onto random walks

● **(s, r, o)**

● **(s, r)**

2. Train ``reasoning" autoregressive models on **2nd half of random walks** (the half encoding **o**)

● reasoning



	HSN$_{(50,\ 20)}$	HSN$^{\text{AR}}_{(50,\ 20)}$	COMET$_{\text{(GPT2)}}$	COMET$_{\text{(GPT2-XL)}}$	COMET $_{\text{(BART)}}$
BLEU-2	**0.462**	0.129	0.225	0.300	0.330
BERT Score	**0.694**	0.374	0.385	0.638	0.650

Object generation quality. COMET$_{\text{(GPT2-XL)}}$ and COMET $_{\text{(BART)}}$ results were extracted from Hwang et al. (AAAI 2021). COMET$_{\text{(GPT2)}}$ was computed by us. All models use greedy decoding for *all* text prefixes in the dataset.