Deriving Language Models from Masked Language Models

Lucas Torroba Hennigen **Yoon Kim**
Massachusetts Institute of Technology
Computer Science and Artificial Intelligence Laboratory
lucastor@mit.edu yoonkim@mit.edu

Abstract

Masked language models (MLM) do not explicitly define a distribution over language, i.e., they are not language models *per se*. However, recent work has implicitly treated them as such for the purposes of generation and scoring. This paper studies methods for deriving explicit joint distributions from MLMs, focusing on distributions over two tokens, which makes it possible to calculate exact distributional properties. We find that an approach based on identifying joints whose conditionals are closest to those of the MLM works well and outperforms existing Markov random field-based approaches. We further find that this derived model's conditionals can even occasionally outperform the original MLM's conditionals.

1 Introduction

Masked language modeling has proven to be an effective paradigm for representation learning (Devlin et al., 2019; Liu et al., 2019; He et al., 2021). However, unlike regular language models, masked language models (MLM) do not define an explicit joint distribution over language. While this is not a serious limitation from a representation learning standpoint, having explicit access to joint distributions would be useful for the purposes of generation (Ghazvininejad et al., 2019), scoring (Salazar et al., 2020), and would moreover enable evaluation of MLMs on standard metrics such as perplexity.

Strictly speaking, MLMs *do* define a joint distribution over tokens that have been masked out. But they assume that the masked tokens are conditionally independent given the unmasked tokens—an assumption that clearly does not hold for language. How might we derive a language model from an MLM such that it does not make unrealistic independence assumptions? One approach is to use the set of the MLM's *unary conditionals*—the conditionals that result from masking just a single token in the input—to construct a fully-connected Markov random field (MRF) over the input (Wang and Cho, 2019; Goyal et al., 2022). This resulting MRF no longer makes any independence assumptions. It is unclear, however, if this heuristic approach actually results in a good language model.[1]

This paper adopts an alternative approach which stems from interpreting the unary conditionals of the MLM as defining a *dependency network* (Heckerman et al., 2000; Yamakoshi et al., 2022).[2] Dependency networks specify the statistical relationship among variables of interest through the set of conditional distributions over each variable given its Markov blanket, which in the MLM case corresponds to all the other tokens. If the conditionals from a dependency network are *compatible*, i.e., there exists a joint distribution whose conditionals coincide with those of the dependency network's, then one can recover said joint using the Hammersley–Clifford–Besag (HCB; Besag, 1974) theorem. If the conditionals are incompatible, then we can adapt approaches from statistics for deriving near-compatible joint distributions from incompatible conditionals (AG; Arnold and Gokhale, 1998).

While these methods give statistically-principled approaches to deriving explicit joints from the MLM's unary conditionals, they are intractable to apply to derive distributions over full sequences. We thus study a focused setting where it is tractable to compute the joints exactly, viz., the *pairwise language model* setting where we use the MLM's unary conditionals of two tokens to derive a joint

[1] MRFs derived this way are still not language models in the strictest sense (e.g., see Du et al., 2022) because the probabilities of sentences of a given length sum to 1, and hence the sum of probabilities of all strings is infinite (analogous to left-to-right language models trained without an [EOS] token; Chen and Goodman, 1998). This can be remedied by incorporating a distribution over sentence lengths.

[2] Recent work by Yamakoshi et al. (2022) has taken this view, focusing on sampling from the dependency network as a means to *implicitly* characterize the joint distribution of an MLM. Here we focus on an *explicit* characterization of the joint.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1149–1159

over these two tokens (conditioned on all the other tokens). Experiments under this setup reveal that AG method performs best in terms of perplexity, with the the HCB and MRF methods performing similarly. Surprisingly, we also find that the unary conditionals of the near-compatible AG joint occasionally have lower perplexity than the original unary conditionals learnt by the MLM, suggesting that regularizing the conditionals to be compatible may be beneficial insofar as modeling the distribution of language.[3]

2 Joint distributions from MLMs

Let $\mathcal{V}$ be a vocabulary, T be the text length, and $\mathbf{w} \in \mathcal{V}^T$ be an input sentence or paragraph. We are particularly interested in the case when a subset $S \subseteq [T] \triangleq \{1, \dots, T\}$ of the input $\mathbf{w}$ is replaced with `[MASK]` tokens; in this case we will use the notation $q_{\{t\}|\overline{S}}(\cdot \mid \mathbf{w}_{\overline{S}})$ to denote the output distribution of the MLM at position $t \in S$, where we mask out the positions in S, i.e., for all $k \in S$ we modify $\mathbf{w}$ by setting $w_k = $ `[MASK]`. If $S = \{t\}$, then we call $q_{t|\overline{t}} \triangleq q_{\{t\}|\overline{\{t\}}}$ a *unary conditional*. Our goal is to use these conditionals to construct joint distributions $q_{S|\overline{S}}(\cdot \mid \mathbf{w}_{\overline{S}})$ for any S.

Direct MLM construction. The simplest approach is to simply mask out the tokens over which we want a joint distribution, and define it to be the product of the MLM conditionals,

$$q^{\text{MLM}}_{S|\overline{S}}(\mathbf{w}_S \mid \mathbf{w}_{\overline{S}}) \triangleq \prod_{i \in S} q_{\{i\}|\overline{S}}(w_i \mid \mathbf{w}_{\overline{S}}). \quad (1)$$

This joint assumes that the entries of $\mathbf{w}_S$ are conditionally independent given $\mathbf{w}_{\overline{S}}$. Since one can show that MLM training is equivalent to learning the conditional marginals of language (App. A), this can be seen as approximating conditionals with a (mean field-like) factorizable distribution.

MRF construction. To address the conditional independence limitation of MLMs, prior work (Wang and Cho, 2019; Goyal et al., 2022) has proposed deriving joints by defining an MRF using the unary conditionals of the MLM. Accordingly, we define

$$q^{\text{MRF}}_{S|\overline{S}}(\mathbf{w}_S \mid \mathbf{w}_{\overline{S}}) \propto \prod_{t \in S} q_{t|\overline{t}}(w_t \mid \mathbf{w}_{\overline{t}}), \quad (2)$$

which can be interpreted as a fully connected MRF, whose log potential is given by the sum of the unary log probabilities. One can similarly define a variant of this MRF where the log potential is the sum of the unary *logits*. MRFs defined this way have a single fully connected clique and thus do not make any conditional independence assumptions. However, such MRFs can have unary conditionals that deviate from the MLM's unary conditionals even if those are compatible (App. B). This is potentially undesirable since the MLM unary conditionals could be close to the true unary conditionals,[4] which means the MRF construction could be worse than the original MLM in terms of unary perplexity.

Hammersley–Clifford–Besag construction. The Hammersley–Clifford–Besag theorem (HCB; Besag, 1974) provides a way of reconstructing a joint distribution from its unary conditionals. Without loss of generality, assume that $S = \{1, \dots, k\}$ for some $k \leq T$. Then given a *pivot point* $\mathbf{w}' = (w'_1, \dots, w'_k) \in \mathcal{V}^k$, we define

$$q^{\text{HCB}}_{S|\overline{S}}(\mathbf{w}_S \mid \mathbf{w}_{\overline{S}}) \propto \prod_{t \in S} \frac{q_{t|\overline{t}}(w_t \mid \mathbf{w}_{>t}, \mathbf{w}'_{<t})}{q_{t|\overline{t}}(w'_t \mid \mathbf{w}_{>t}, \mathbf{w}'_{<t})}, \quad (3)$$

where $\mathbf{w}'_{<i} \triangleq (w'_1, \dots, w'_{i-1})$, and similarly $\mathbf{w}_{>i} \triangleq (w_{i+1}, \dots, w_T)$. Importantly, unlike the MRF approach, if the unary conditionals of the MLM *are* compatible, then HCB will recover the true joint, irrespective of the choice of pivot.

Arnold–Gokhale construction. If we assume that the unary conditionals are not compatible, then we can frame our goal as finding a near-compatible joint, i.e., a joint such that its unary conditionals are close to the unary conditionals of the MLM. Formally, for any S and fixed inputs $\mathbf{w}_{\overline{S}}$, we can define this objective as,

$$q^{\text{AG}}_{S|\overline{S}}(\cdot \mid \mathbf{w}_{\overline{S}}) = \operatorname*{argmin}_{\mu} \sum_{t \in S} \sum_{\mathbf{w}' \in \mathcal{V}^{|S|-1}} J(t, \mathbf{w}'), \quad (4)$$

where $J(t, \mathbf{w}')$ is defined as:

$$\text{KL}(q_{t|S\setminus\{t\},\overline{S}}(\cdot \mid \mathbf{w}', \mathbf{w}_{\overline{S}}) \,||\, \mu_{t|S\setminus\{t\},\overline{S}}(\cdot \mid \mathbf{w}', \mathbf{w}_{\overline{S}})).$$

We can solve this optimization problem using Arnold and Gokhale's (1998) algorithm (App. C).

2.1 Pairwise language model

In language modeling we are typically interested in the probability of a sequence $p(\mathbf{w})$. However, the above methods are intractable to apply to full sequences (except for the baseline MLM). For example, the lack of any independence assumptions

[3] Our code and data is available at: https://github.com/ltorroba/lms-from-mlms.

[4] As noted by https://machinethoughts.wordpress.com/2019/07/14/a-consistency-theorem-for-bert/

in the MRF means that the partition function requires full enumeration over V^T sequences.[5] We thus focus our empirical study on the pairwise setting where $|S| = 2$.[6] In this setting, we can calculate $q_{S|\overline{S}}(\cdot \mid \mathbf{w}_{\overline{S}})$ with $O(V)$ forward passes of the MLM for all methods.

3 Evaluation

We compute two sets of metrics that evaluate the resulting joints in terms of (i) how good they are as probabilistic models of language and (ii) how faithful they are to the original MLM conditionals (which are trained to approximate the true conditionals of language, see App. A). Let $\mathcal{D} = \{(\mathbf{w}^{(n)}, S^{(n)})\}_{n=1}^{N}$ be a dataset where $\mathbf{w}^{(n)}$ is an English sentence and $S^{(n)} = (a^{(n)}, b^{(n)})$ are the two positions being masked. We define the following metrics to evaluate a distribution q':

Language model performance. We consider two performance metrics. The first is the pairwise perplexity (**P-PPL**) over two tokens,

$$\exp\left(\frac{-1}{2N}\sum_{n=1}^{N} \log q'_{a^{(n)},b^{(n)}|\overline{S}^{(n)}}(w_{a^{(n)}}^{(n)}, w_{b^{(n)}}^{(n)} \mid \mathbf{w}_{\overline{S}^{(n)}}^{(n)})\right)$$

We would expect a good joint to obtain lower pairwise perplexity than the original MLM, which (wrongly) assumes conditional independence. The second is unary perplexity (**U-PPL**),

$$\exp\left(\frac{-1}{2N}\sum_{n=1}^{N} \sum_{\substack{(i,j)\in \\ \{S^{(n)},S_r^{(n)}\}}} \log q'_{i|j,\overline{S}^{(n)}}(w_i^{(n)} \mid w_j^{(n)}, \mathbf{w}_{\overline{S}^{(n)}}^{(n)})\right)$$

where for convenience we let $S_r^{(n)} \triangleq (b^{(n)}, a^{(n)})$ as the reverse of the masked positions tuple $S^{(n)}$. Note that this metric uses the unary conditionals derived from the pairwise joint, i.e., $q'_{i|j,S}$, except in the MLM construction case which uses the MLM's original unary conditionals.

Faithfulness. We also assess how faithful the new unary conditionals are to the original unary conditionals by calculating the average conditional KL divergence (**A-KL**) between them,

$$\sum_{n=1}^{N}\sum_{w'\in\mathcal{V}} \frac{\mathrm{D}(S^{(n)}, w', \mathbf{w}_{\overline{S}^{(n)}}^{(n)}) + \mathrm{D}(S_r^{(n)}, w', \mathbf{w}_{\overline{S}^{(n)}}^{(n)})}{2N|\mathcal{V}|}.$$

where we define $\mathrm{D}(S, w', \mathbf{w}_{\overline{S}}) \triangleq \mathrm{KL}(q_{a|b,\overline{S}}(\cdot \mid w', \mathbf{w}_{\overline{S}}) \,||\, q'_{a|b,\overline{S}}(\cdot \mid w', \mathbf{w}_{\overline{S}}))$ for $S = (a, b)$. If the new joint is completely faithful to the MLM, this number should be zero. The above metric averages the KL across the entire vocabulary $\mathcal{V}$, but in practice we may be interested in assessing closeness only when conditioned on the gold tokens. We thus compute a variant of the above metric where we only average over the conditionals for the gold token (**G-KL**):

$$\sum_{n=1}^{N} \frac{\mathrm{D}(S^{(n)}, w_{b^{(n)}}^{(n)}, \mathbf{w}_{\overline{S}^{(n)}}^{(n)}) + \mathrm{D}(S_r^{(n)}, w_{a^{(n)}}^{(n)}, \mathbf{w}_{\overline{S}^{(n)}}^{(n)})}{2N}.$$

This metric penalizes unfaithfulness in common contexts more than in uncommon contexts. Note that if the MLM's unary conditionals are compatible, then both the HCB and AG approach should yield the same joint distribution, and their faithfulness metrics should be zero.

3.1 Experimental setup

We calculate the above metrics on 1000 examples[7] from a natural language inference dataset (SNLI; Bowman et al., 2015) and a summarization dataset (XSUM; Narayan et al., 2018). We consider two schemes for selecting the tokens to be masked for each sentence: masks over two tokens chosen uniformly at random (**Random pairs**), and also over random *contiguous* tokens in a sentence (**Contiguous pairs**). Since inter-token dependencies are more likely to emerge when adjacent tokens are masked, the contiguous setup magnifies the importance of deriving a good pairwise joint. In addition, we consider both $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ (cased) as the MLMs from which to obtain the unary conditionals.[8] For the AG joint, we run $t = 50$ steps of Arnold and Gokhale's (1998) algorithm (App. C), which was enough for convergence. For the HCB joint, we pick a pivot using the mode of the pairwise joint of the MLM.[9]

4 Results

The results are shown in Tab. 1. Comparing the PPL's of MRF and MRF_{L} (i.e., the MRF using logits), the former consistently outperforms the latter,

[5] We also tried estimating the partition through importance sampling with GPT-2 but found the estimate to be quite poor.

[6] Concurrent work by Young and You (2023) also explores the (in)compatibility of MLMs in the $|S| = 2$ case.

[7] Each example requires running the MLM over 28 000 times, so it is expensive to evaluate on many more examples.

[8] Specifically, we use the bert-base-cased and bert-large-cased implementations from HuggingFace (Wolf et al., 2020).

[9] We did not find HCB to be too sensitive to the pivot in preliminary experiments.

	Dataset	Scheme	Random pairs U-PPL	Random pairs P-PPL	Random pairs A-KL	Random pairs G-KL	Contiguous pairs U-PPL	Contiguous pairs P-PPL	Contiguous pairs A-KL	Contiguous pairs G-KL
Ⓑ	SNLI	MLM	11.22	19.01	1.080	0.547	13.78	74.68	4.014	1.876
		$\mathrm{MRF_L}$	13.39	71.44	0.433	0.267	23.45	13 568.17	1.543	0.607
		MRF	12.30	21.65	0.658	0.179	18.35	126.05	1.967	0.366
		HCB	12.51	22.62	0.593	0.168	17.71	589.02	2.099	0.416
		AG	10.76	12.68	0.007	0.085	13.26	21.59	0.018	0.181
	XSUM	MLM	4.88	6.12	0.404	0.227	4.91	39.33	4.381	2.128
		$\mathrm{MRF_L}$	5.17	9.12	0.148	0.085	6.55	2209.94	1.561	0.383
		MRF	5.00	6.23	0.262	0.049	5.53	47.62	2.242	0.185
		HCB	5.08	6.21	0.256	0.052	6.46	174.32	2.681	0.328
		AG	5.00	5.29	0.003	0.044	5.27	8.42	0.016	0.143
Ⓛ	SNLI	MLM	9.50	18.57	1.374	0.787	10.42	104.12	4.582	2.463
		$\mathrm{MRF_L}$	11.52	76.23	0.449	0.276	15.43	8536.92	1.470	0.543
		MRF	10.57	19.54	0.723	0.193	13.07	93.33	1.992	0.359
		HCB	10.71	20.70	0.797	0.215	14.43	458.25	2.563	0.552
		AG	8.57	10.11	0.007	0.097	9.64	15.64	0.019	0.173
	XSUM	MLM	3.80	5.67	0.530	0.413	3.91	103.86	5.046	3.276
		$\mathrm{MRF_L}$	3.94	7.06	0.156	0.068	4.62	1328.20	1.441	0.290
		MRF	3.87	4.94	0.322	0.036	4.16	36.66	2.258	0.145
		HCB	3.91	5.14	0.346	0.059	5.67	164.15	2.954	0.400
		AG	3.88	4.13	0.003	0.042	4.21	6.62	0.016	0.126

Table 1: Comparison of MRF, HCB and AG constructions on randomly sampled SNLI (Bowman et al., 2015) sentences and XSUM (Narayan et al., 2018) summaries. We apply the constructions to two MLMs: $\mathrm{BERT_{BASE}}$ (Ⓑ) and $\mathrm{BERT_{LARGE}}$ (Ⓛ). We consider both masking tokens uniformly at random (Random pairs) and masking adjacent tokens uniformly at random (Contiguous pairs). For all metrics, lower is better.

indicating that using the raw logits generally results in a worse language model. Comparing the MRFs to MLM, we see that the unary perplexity (U-PPL) of the MLM is lower than those of the MRFs, and that the difference is most pronounced in the contiguous masking case. More surprisingly, we see that the pairwise perplexity (P-PPL) is often (much) higher than the MLM's, even though the MLM makes unrealistic conditional independence assumptions. These results suggest that the derived MRFs are in general worse unary/pairwise probabilistic models of language than the MLM itself, implying that the MRF heuristic is inadequate (see App. D for a qualitative example illustrating how this can happen). Finally, we also find that the MRFs' unary conditionals are not faithful to those of the MRFs based on the KL measures. Since one can show that the MRF construction can have unary conditionals that have nonzero KL to the MLM's unary conditionals even if they are compatible (App. B), this gives both theoretical and empirical arguments against the MRF construction.

The HCB joint obtains comparable performance to MRF in the random masking case. In the contiguous case, it exhibits similar failure modes as the MRF in producing extremely high pairwise perplexity (P-PPL) values. The faithfulness metrics are similar to the MRF's, which suggests that the conditionals learnt by MLMs are incompatible. The AG approach, on the other hand, outperforms the $\mathrm{MRF_L}$, MRF and HCB approaches in virtually all metrics. This is most evident in the contiguous masking case, where AG attains lower pairwise perplexity than all models, including the MLM itself. In some cases, we find that the AG model even outperforms the MLM in terms of unary perplexity, which is remarkable since the unary conditionals of the MLM were *trained* to approximate the unary conditionals of language (App. A). This indicates that near-compatibility may have regularizing effect that leads to improved MLMs. Since AG was optimized to be near-compatible, its joints are unsurprisingly much more faithful to the original MLM's conditionals. However, AG's G-KL tends to be on par with the other models, which suggests that it is still not faithful to the MLM in the contexts that are most likely to arise. Finally, we analyze the effect of masked position distance on language modeling performance, and find that improvements are most pronounced when the masked tokens are close to each other (see App. E).

5 Related work

Probabilistic interpretations of MLMs. In one of the earliest works about sampling from MLMs, Wang and Cho (2019) propose to use unary conditionals to sample sentences. Recently Yamakoshi et al. (2022) highlight that, while this approach only constitutes a pseudo-Gibbs sampler, the act of re-sampling positions uniformly at random guarantees that the resulting Markov chain has a unique, stationary distribution (Bengio et al., 2013, 2014). Alternatively, Goyal et al. (2022) propose defining an MRF from the MLM's unary conditionals, and sample from this via Metropolis-Hastings. Concurrently, Young and You (2023) conduct an empirical study of the compatibility of BERT's conditionals.

Compatible distributions. The statistics community has long studied the problem of assessing the compatibility of a set of conditionals (Arnold and Press, 1989; Gelman and Speed, 1993; Wang and Kuo, 2010; Song et al., 2010). Arnold and Gokhale (1998) and Arnold et al. (2002) explore algorithms for reconstructing near-compatible joints from incompatible conditionals, which we leverage in our work. Besag (1974) also explores this problem, and defines a procedure (viz., eq. 3) for doing so when the joint distribution is strictly positive and the conditionals are compatible. Lowd (2012) apply a version of HCB to derive Markov networks from incompatible dependency networks (Heckerman et al., 2000).

6 Conclusion

In this paper, we studied four different methods for deriving an explicit joint distributions from MLMs, focusing in the pairwise language model setting where it is possible to compute exact distributional properties. We find that the Arnold–Gokhale (AG) approach, which finds a joint whose conditionals are closest to the unary conditionals of an MLM, works best. Indeed, our results indicate that said conditionals can attain lower perplexity than the unary conditionals of the original MLM. It would be interesting to explore whether explicitly regularizing the conditionals to be compatible during MLM training would lead to better modeling of the distribution of language.

7 Limitations

Our study illuminates the deficiencies of the MRF approach and applies statistically-motivated approaches to craft more performant probabilistic models. However, it is admittedly not clear how these insights can immediately be applied to improve downstream NLP tasks. We focused on models over pairwise tokens in order to avoid sampling and work with exact distributions for the various approaches (MRF, HCB, AG). However this limits the generality of our approach (e.g., we cannot score full sentences). We nonetheless believe that our empirical study is interesting on its own and suggests new paths for developing efficient and faithful MLMs.

Ethics Statement

We foresee no ethical concerns with this work.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This research is supported in part by funds from the MLA@CSAIL initiative and MIT-IBM Watson AI lab. LTH acknowledges support from the Michael Athans fellowship fund.

References

Barry C. Arnold, Enrique Castillo, and José María Sarabia. 2002. Exact and near compatibility of discrete conditional distributions. *Computational Statistics & Data Analysis*, 40(2):231–252.

Barry C. Arnold and Dattaprabhakar V. Gokhale. 1998. Distributions most nearly compatible with given families of conditional distributions. *Test*, 7(2):377–390.

Barry C. Arnold and James S. Press. 1989. Compatible conditional distributions. *Journal of the American Statistical Association*, 84(405):152–156.

Yoshua Bengio, Éric Thibodeau-Laufer, Guillaume Alain, and Jason Yosinski. 2014. Deep generative stochastic networks trainable by backprop. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 226–234, Bejing, China. PMLR.

Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. 2013. Generalized denoising auto-encoders as generative models. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS, page 899–907, Red Hook, New York, USA. Curran Associates Inc.

Julian Besag. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, 36(2):192–236.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical report, Harvard University.

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, USA. Association for Computational Linguistics.

Li Du, Lucas Torroba Hennigen, Tiago Pimentel, Clara Meister, Jason Eisner, and Ryan Cotterell. 2022. A measure-theoretic characterization of tight language models.

Andrew Gelman and Terence P. Speed. 1993. Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society*, 55(1):185–188.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-Predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6112–6121, Hong Kong, China. 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 *International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.

David Heckerman, Max Chickering, Chris Meek, Robert Rounthwaite, and Carl Kadie. 2000. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1:49–75.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*.

Daniel Lowd. 2012. Closed-form learning of markov networks from dependency networks. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 533–542, Catalina Island, California, USA. Association for Uncertainity in Artificial Intelligence.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Chwan-Chin Song, Lung-An Li, Chong-Hong Chen, Thomas J. Jiang, and Kun-Lin Kuo. 2010. Compatibility of finite discrete conditional distributions. *Statistica Sinica*, 20(1):423–440.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yuchung J. Wang and Kun-Lin Kuo. 2010. Compatibility of discrete conditional distributions with structural zeros. *Journal of Multivariate Analysis*, 101(1):191–199.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Takateru Yamakoshi, Thomas Griffiths, and Robert Hawkins. 2022. Probing BERT's priors with serial reproduction chains. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3977–3992, Dublin, Ireland. Association for Computational Linguistics.

Tom Young and Yang You. 2023. On the inconsistencies of conditionals learned by masked language models.

A MLMs as learning conditional marginals

One can show that the MLM training objective corresponds to learning to approximate the conditional marginals of language, i.e., the (single-position) marginals of language when we condition on any particular set of positions. More formally, consider an MLM parameterized by a vector $\boldsymbol{\theta} \in \Theta$ and some distribution $\mu(\cdot)$ over positions to mask $S \subseteq [T]$. Then the MLM learning objective is given by:

$$\hat{\boldsymbol{\theta}} = \operatorname*{argsup}_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{S\sim\mu(\cdot)} \mathop{\mathbb{E}}_{\mathbf{w}\sim p(\cdot)} \left[\frac{1}{|S|} \sum_{t\in S} \log q_{t|\overline{S}}(w_t \mid \mathbf{w}_{\overline{S}}; \boldsymbol{\theta}) \right],$$

where $p(\cdot)$ denotes the true data distribution. Analogously, let $p_{S|\overline{S}}(\cdot \mid \mathbf{w}_{\overline{S}})$ and $p_{\overline{S}}(\cdot)$ denote the conditionals and marginals of the data distribution, respectively. Then the above can be rewritten as:

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \operatorname*{argsup}_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{S\sim\mu(\cdot)} \mathop{\mathbb{E}}_{\mathbf{w}_{\overline{S}}\sim p_{\overline{S}}(\cdot)} \left[\frac{1}{|S|} \sum_{t\in S} \mathop{\mathbb{E}}_{\mathbf{w}_S\sim p_{S|\overline{S}}(\cdot)} \left[\log q_{t|\overline{S}}(w_t \mid \mathbf{w}_{\overline{S}}; \boldsymbol{\theta}) \right] \right] \\
&= \operatorname*{arginf}_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{S\sim\mu(\cdot)} \mathop{\mathbb{E}}_{\mathbf{w}_{\overline{S}}\sim p_{\overline{S}}(\cdot)} \left[\frac{1}{|S|} \sum_{t\in S} \mathrm{KL}(p_{t|\overline{S}}(\cdot \mid \mathbf{w}_{\overline{S}}) \,||\, q_{t|\overline{S}}(\cdot \mid \mathbf{w}_{\overline{S}}; \boldsymbol{\theta})) \right],
\end{aligned}$$

Thus, we can interpret MLM training as learning to approximate the conditional marginals of language, i.e., $\forall S \subseteq [T]$ and $\forall t \in S$, in the limit we would expect that, for any observed context $\mathbf{w}_{\overline{S}}$, we have $q_{t|\overline{S}}(\cdot \mid \mathbf{w}_{\overline{S}}) \approx p_{t|\overline{S}}(\cdot \mid \mathbf{w}_{\overline{S}})$.

B Unfaithful MRFs

Here we show that even if the unary conditionals used in the MRF construction are compatible (Arnold and Press, 1989), the unary conditionals of the probabilistic model implied by the MRF construction can deviate (in the KL sense) from the true conditionals. This is important because (i) it suggests that we might do better (at least in terms of U-PPL) by simply sticking to the conditionals learned by MLM, and (ii) this is not the case for either the HCB or the AG constructions, i.e., if we started with the correct conditionals, HCB and AG's joint would be compatible with the MLM. Formally,

Proposition B.1. Let $w_1, w_2 \in \mathcal{V}$ and further let $p_{1|2}(\cdot \mid w_2), p_{2|1}(\cdot \mid w_1)$ be the true (i.e., population) unary conditional distributions. Define an MRF as

$$q_{1,2}(w_1, w_2) \propto p_{1|2}(w_1 \mid w_2)\, p_{2|1}(w_2 \mid w_1),$$

and let $q_{1|2}(\cdot \mid w_2), q_{2|1}(\cdot \mid w_1)$ be the conditionals derived from the MRF. Then there exists $p_{1|2}, p_{2|1}$ such that

$$\mathrm{KL}(p_{1|2}(\cdot \mid w_2) \,||\, q_{1|2}(\cdot \mid w_2)) > 0.$$

Proof. Let $w_2 \in \mathcal{V}$ be arbitrary. We then have:

$$q_{1|2}(w_1 \mid w_2) = \frac{p_{1|2}(w_1 \mid w_2)\, p_{2|1}(w_2 \mid w_1)}{\sum_{w'\in\mathcal{V}} p_{1|2}(w' \mid w_2)\, p_{2|1}(w_2 \mid w')}$$

Now, consider the KL between the true unary conditionals and the MRF unary conditionals:

$$\begin{aligned}
\mathrm{KL}(p_{1|2}(\cdot \mid w_2) \,||\, q_{1|2}(\cdot \mid w_2)) &= \sum_{w\in\mathcal{V}} p_{1|2}(w \mid w_2) \log \frac{p_{1|2}(w \mid w_2)}{q_{1|2}(w \mid w_2)} \\
&= \sum_{w\in\mathcal{V}} p_{1|2}(w \mid w_2) \log \frac{\sum_{w'\in\mathcal{V}} p_{1|2}(w' \mid w_2)\, p_{2|1}(w_2 \mid w')}{p_{2|1}(w_2 \mid w)} \\
&= \log \mathbb{E}_{w\sim p_{1|2}(\cdot|w_2)}[p_{2|1}(w_2 \mid w)] - \mathbb{E}_{w\sim p_{1|2}(\cdot|w_2)}[\log p_{2|1}(w_2 \mid w)]
\end{aligned}$$

This term is the Jensen gap, and in general it can be non-zero. To see this, suppose $\mathcal{V} = \{a, b\}$ and consider the joint

$$p_{1,2}(w_1, w_2) = \begin{cases} \frac{97}{100} & w_1, w_2 = a \\ \frac{1}{100} & \text{otherwise} \end{cases}$$

with corresponding conditionals $p_{2|1}(x \mid b) = p_{1|2}(x \mid b) = \frac{1}{2}$ for all $x \in \mathcal{V}$ and

$$p_{2|1}(x \mid a) = p_{1|2}(x \mid a) = \begin{cases} \frac{97}{98} & x = a \\ \frac{1}{98} & x = b \end{cases}$$

Now, take $w_2 = b$. We then have

$$\begin{aligned} &\mathrm{KL}(p_{1|2}(\cdot \mid b) \,||\, q_{1|2}(\cdot \mid b)) \\ &\quad = \log \mathbb{E}_{w \sim p_{1|2}(\cdot|b)}[p_{2|1}(b \mid w)] - \mathbb{E}_{w \sim p_{1|2}(\cdot|b)}[\log p_{2|1}(b \mid w)] \\ &\quad = \log \left(\frac{1}{2} \left(\frac{1}{98} + \frac{1}{2} \right) \right) - \frac{1}{2} \left(\log \frac{1}{98} + \log \frac{1}{2} \right) \\ &\quad = \log \left(\frac{1}{196} + \frac{1}{4} \right) - \frac{1}{2} \left(\log \frac{1}{196} \right) \approx 1.27 \end{aligned}$$

which demonstrates that the KL can be non-zero.

□

C Arnold–Gokhale algorithm

Arnold and Gokhale (1998) study the problem of finding a near-compatible joint from unary conditionals, and provide and algorithm for the case of $|S| = 2$. The algorithm initializes the starting pairwise distribution $q^{\text{AG(1)}}_{a,b|\overline{S}}(\cdot, \cdot \mid \mathbf{w}_{\overline{S}})$ to be uniform, and performs the following update until convergence:

$$q^{\text{AG(t+1)}}_{a,b|\overline{S}}(w_a, w_b \mid \mathbf{w}_{\overline{S}}) \propto \frac{q_{a|b,\overline{S}}(w_a \mid w_b, \mathbf{w}_{\overline{S}}) + q_{b|a,\overline{S}}(w_b \mid w_a, \mathbf{w}_{\overline{S}})}{\left(q^{\text{AG(t)}}_{a|\overline{S}}(w_a \mid \mathbf{w}_{\overline{S}})\right)^{-1} + \left(q^{\text{AG(t)}}_{b|\overline{S}}(w_b \mid \mathbf{w}_{\overline{S}})\right)^{-1}}. \quad (5)$$

D Qualitative example of MRF underperformance

This example from SNLI qualitatively illustrates a case where both the unary and pairwise perplexities from the MRF underperforms the MLM: "The [MASK]$_1$ [MASK]$_2$ at the casino", where the tokens "man is" are masked. In this case, both MRFs assign virtually zero probability mass to the correct tokens, while the MLM assigns orders of magnitude more (around 0.2% of the mass of the joint). Upon inspection, this arises because $q_{2|1,\overline{S}}(\text{is} \mid \text{man}) \approx 0.02$ and $q_{1|2,\overline{S}}(\text{man} \mid \text{is}) \approx 2 \times 10^{-5}$, which makes the numerator of $q^{\text{MRF}}_{1,2|\overline{S}}(\text{man}, \text{is})$ be ≈ 0. The MRF could still assign high probability to this pair if the denominator is also ≈ 0, but in this case we have $q_{2|1,\overline{S}}(\text{was} \mid \text{man}) \approx 0.33$ and $q_{1|2,\overline{S}}(\text{man} \mid \text{was}) \approx 0.03$, which makes the denominator well above 0. This causes the completion "man is" to have disproportionately little mass in the joint compared other to combinations ("man was") that were ascribed more mass by BERT's unary conditionals.

E Token distance analysis

We also explore the effect of the distance between masked tokens on the pairwise negative log-likelihood (PNLL, lower is better; note this is equivalent to the log PPPL) of the joints built using the different approaches we considered. We considered two different kinds of distance functions between tokens: (i) the absolute difference in the positions between the two masked tokens, and (ii) their syntactic distance (obtained by running a dependency parser on unmasked sentences).

We plot the results in Fig. 1 (SNLI) and Fig. 2 (XSUM). Note that the black bars denote the number of datapoints with that distance between the two masked tokens, where a syntactic distance of 0 means that the two masked tokens belong to the same word, whereas a token distance of 0 means that the two masked tokens are adjacent. The graphs indicate that the language modeling performance improvement (compared to using the MLM joint) is most prominent when masked tokens are close together, which is probably because when the masked tokens are close together they are more likely to be dependent. In this case, AG tends to do best, HCB and MRF tend to do similarly, followed by MRF-L and, finally, the conditionally independent MLM, which follows the trends observed in the paper.

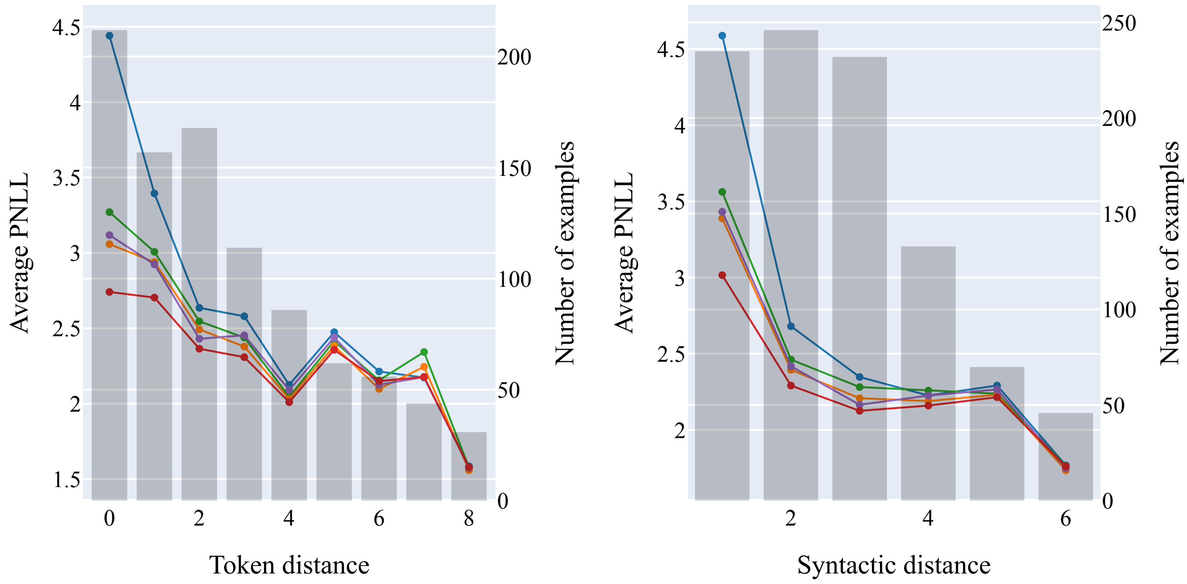

Figure 1: Pairwise NLL (PNLL) as a function of the token and syntactic distance between masked positions for joints built using the methods: MLM, MRF (Logit), MRF, HCB, AG on SNLI (Bowman et al., 2015). The gray bars represent the number of examples on the dataset that had that degree of separation.

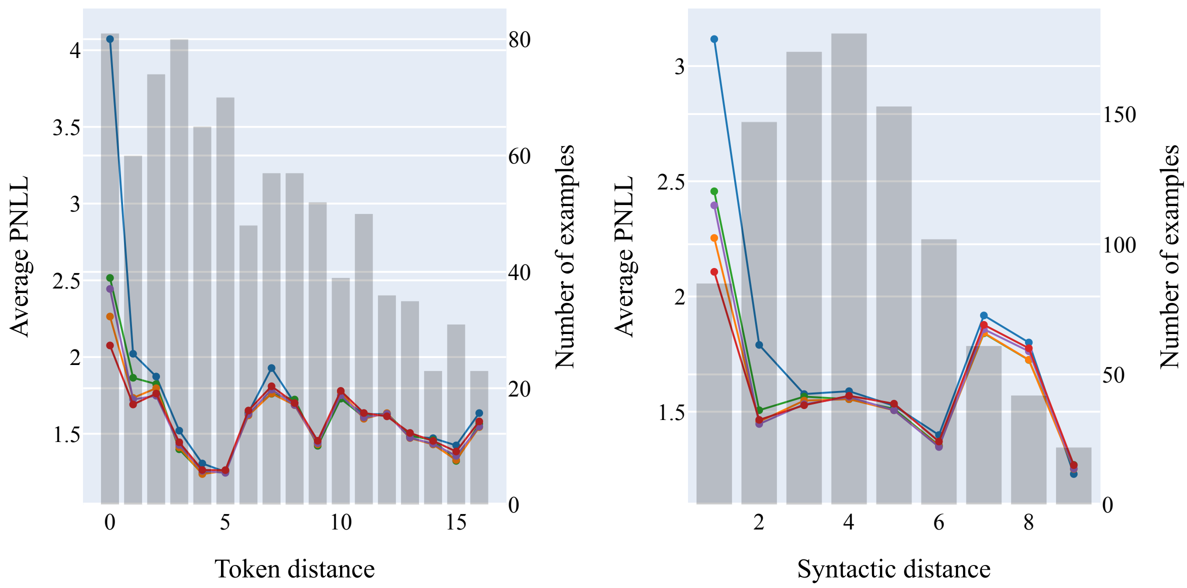

Figure 2: Pairwise NLL (PNLL) as a function of the token and syntactic distance between masked positions for joints built using the methods: MLM, MRF (Logit), MRF, HCB, AG on XSUM (Narayan et al., 2018). The gray bars represent the number of examples on the dataset that had that degree of separation.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6

☑ A2. Did you discuss any potential risks of your work?
Section 7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3

☑ B1. Did you cite the creators of artifacts you used?
Section 3

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Available online

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Consistent with intended use

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3

C ☑ Did you run computational experiments?

Section 4

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Available online

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. Left blank.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.