Instructions for EMNLP 2020 Proceedings

Anonymous EMNLP submission

Abstract

todo

1 Introduction

Hidden Markov models (HMMs) are a fundamental model in machine learning, providing the simplest latent-variable model for sequential data. HMMs are core to time-series problems in bioinformatics, reinforcement learning, and, of course, natural language processing. HMMs have remained a core building block, because they cleanly separate out the facet of discrete model memory from the generation of observations. Historically they have been used extensively in NLP for tasks such as sequence modeling (Rabiner, 1990), alignment (Vogel et al., 1996), and less extensively language modeling (Kuhn et al., 1994; Huang, 2011).

In recent years, most state-of-the-art systems in NLP have moved away utilizing explicit hidden states. (Something about sequence modleing) particularly for structured models. In the area of language models modeled the observed words using n-gram models, feedforward neural networks (Bengio et al., 2003), and eventually recurrent neural networks (Mikolov et al., 2010; Zaremba et al., 2014; Merity et al., 2017) or transformers (Radford et al., 2019). This progress has led to the common wisdom that latent-variable models are not competitive with fully observed models.

In this work, we revisit the question of developing HMMs for natural language processing, in particular looking at the tasks of language modeling and sequence tagging. We approach this question by taking direct lessons from the success of deep neural models for NLP tasks. In particular we make three observations: (1) The right parameterization is critically important for representation learning, e.g. a feedforward model Bengio et al. (2003) can have the same probabilistic structure as an n-gram model while performing significantly better; (2) overparameterization seems to be critical for finding better optima, e.g. large LSTMs Zaremba et al. (2014) show marked improvements in performance; (3) regularization choices are necessary to handle reconcile (different) parameterizations, e.g. experiments by Merity et al. (2017) outline a variety of training choices.

These observations motivate a new approach to classical HMMs. We develop a neural parameterization for HMMs that extends them to comparable size and structure of deep learning models, while allowing us to lazily instantiate distributions. We also develop a sparsity constraint that allows us to utilize very large HMMs, significantly beyond standard efficiency limits. Finally we incorporate a variant of dropout that both improves accuracy and reduces the computational overhead by an order of magnitude during training.

Experiments show that on two language modeling datasets and a supervised sequence tagging task. We find that our HMM extension significantly outperforms past HMMs as well as n-gram models. It also performs comparably to neural counterparts with a similar number of parameters while maintaining uncertainty over the state dynamics.

2 Related work

Prior work has demonstrated the benefits of neural parameterization of structured generative models. For HMMs, Tran et al. (2016) demonstrated improvements in POS induction with a neural parameterization of an HMM, while Kim et al. (2019) demonstrated improvements in grammar induction with a probabilistic context free grammar. Both of these works used latent variables with relatively small state spaces, as the goal of both was structure

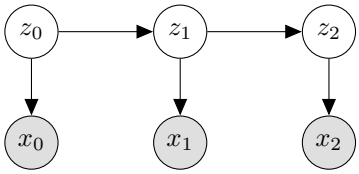

Figure 1: An HMM with tokens x_t and states z_t.

induction rather than language modeling itself. We extend the neural parameterization to much larger state space models.

We also draw inspiration from the experiments with cloned HMMs by Dedieu et al. (2019), who proposed to introduce sparsity constraints in scalar emission distribution of HMMs in order to make conditional inference tractable in large state spaces. Dedieu et al. (2019) trained a 30k state HMM on character-level language modeling by constraining every state to emit only a single character type. This particular constraint is problematic for language modeling at the word level, where the vocabulary size is much larger. We build on their work by proposing a sparsity constraint based on Brown clustering (Brown et al., 1992) which allows us to extend their work to vocabularies that are larger than the state space.

Another approach to scaling to larger state spaces is to initialize with a small state space then grow the state space via a split-merge process (Petrov et al., 2006; Huang, 2011). In particular, Huang (2011) learn an HMM for language modeling via this process. Additionally, the cloned HMM (Dedieu et al., 2019) can be seen as an HMM that starts with a single state per word, then splits every state into k states at the start with no subsequent splits or merges. The application of split-merge is an avenue for future work, as we focus on fixed-size state spaces for simplicity.

Other investigations into improving the performance of HMMs involved relaxing independence or modeling assumptions (Buys et al., 2018) to resemble a recurrent neural network, or through model combination with a recurrent neural network (Krakovna and Doshi-Velez, 2016). There are also extensions of HMMs, such as factorial HMMs (Ghahramani and Jordan, 1997; Nepal and Yates, 2013) and context free grammars (Kim et al., 2019). We leave scaling more expressive models to large state spaces for future work, and focus on scaling the basic HMM.

3 Background: HMMs

We are interested in learning a distribution over observed tokens $\mathbf{x} = \langle x_1, \ldots, x_T \rangle$, with each token x_t an element of the finite vocabulary $\mathcal{X}$. Hidden Markov models (HMMs) specify a joint distribution over observed tokens $\mathbf{x}$ and discrete latent states $\mathbf{z} = \langle z_1, \ldots, z_T \rangle$, with each z_t from finite set $\mathcal{Z}$. For notational convenience, we define the starting state $z_0 = \epsilon$. The model is defined by the following generative process (shown in figure 1): For every time step $t \in \{0, \ldots, T\}$, choose a state given the previous state $z_t \mid z_{t-1}$ from the transition distribution $p(z_t \mid z_{t-1})$. Then choose a token given the current state $x_t \mid z_t$ from the emission distribution $p(x_t \mid z_t)$. This yields the joint distribution

$$p(\mathbf{x}, \mathbf{z}; \theta) = \prod_{t=1}^{T} p(x_t \mid z_t) p(z_t \mid z_{t-1}) \quad (1)$$

The distributions are parameterized as follows

$$\begin{aligned} p(z_1 | z_0) &\propto e^{\psi_{z_1}} \\ p(z_t \mid z_{t-1}) &\propto e^{\psi_{z_t z_{t-1}}} \\ p(x_t \mid z_t) &\propto e^{\phi_{x_t z_t}} \end{aligned} \quad (2)$$

where each $\psi_{z_0}, \psi_{z_t z_{t-1}}, \phi_{x_t z_t} \in \mathbb{R}$. Thus the transition distribution $p(z_t \mid z_{t-1})$ is parameterized by $\psi \in \mathbb{R}^{|\mathcal{Z}| \times |\mathcal{Z}|}$, and the emission distribution $p(x_t \mid z_t)$ by $\phi \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{Z}|}$.

We distinguish two types of parameterizations: *scalar* and *neural*. A scalar parameterization simply uses $\theta = \{\phi, \psi\}$ to fit one model parameter for each distributional parameter. This results in $O(|\mathcal{Z}|^2 + |\mathcal{X}||\mathcal{Z}|)$ model parameters, since the transition and emission matrices must be explicitly represented. This parameterization can lead to either in over or under parameterization for most NLP tasks. In contrast a neural parameterization uses θ as the parameters of a neural network that generates ϕ and ψ for the distribution. These may be customized for the task through embedding or other layers, and separate the parameterization size from the hidden state size.

In order to fit an HMM to data $\mathbf{x}$, we must marginalize over the latent states to obtain the likelihood $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z})$. This sum can be computed in time $O(T|\mathcal{Z}|^2)$ via dynamic programming, which becomes prohibitive if the number of latent states $|\mathcal{Z}|$ is large. As the dynamic program is differentiable, we can optimize the likelihood with

gradient ascent. An alternative method would be to employ expectation maximization, however there is no closed form solution for the M-step with a neural parameterization. Gradient ascent is applicable to both the scalar and neural parameterizations.

(Talk about constructing matrices once per batch, decoupling cost of softmax over vocab and batch size, constructing log potentials via indexing, then logarithmic reduction DP.)

Comparison to RNNs Whereas RNNs encode the previous observations with a distributed representation, HMMs encode the past as a distribution over a finite set of states. We refer to both of the latent representations in the RNN and HMM as belief states, which get updated after observing each token x_t. As noted by Buys et al. (2018), the belief states of the RNN and HMM are used in similar ways. The computational complexity of updating the belief state is equivalent to the belief state update in an HMM if the dimension of the RNN is equal to the number of states in the HMM.

Both RNN and HMM belief updates involve a matrix-vector product, which is the dominating factor computationally. RNNs use a linear projection of previous hidden vectors, while HMMs consider transitions from every previous state to all current states. A key distinguishing characteristic of the RNN belief update is that it contains nonlinearities, resulting in a non-associative operator. On the other hand, the HMMs belief update is associative, allowing us to take advantage of parallel hardware to perform marginal inference in time logarithmic in the length of a sequence. This is similar to Quasi-RNNs (Bradbury et al., 2016), which also have an associative belief update operator albeit without maintaining any notion of uncertainty over states.

(Can expand on this if necessary, or put in appendix. this is hard to talk about without giving the forward algorithm for computing $p(x_t \mid \mathbf{x}_{<t})$)

(seems like this went off the rails, need to simplify / break up)

4 Methods

HMMs cleanly separate out latent memory from the generative process of language. This separation makes them an important model for isolating observed properties and performing inference. However, this separation also means they require a large hidden state space to capture the underlying properties of language.

In this section, we consider three methods for greatly expanding the state space of HMMs without similarly increasing inference complexity or parameterization size.

4.1 Emission Sparsity Constraints

Marginal inference for general HMMs is quadratic in the state space $|\mathcal{Z}|$. In practice, this limits the size of the state space to the order of 100s, which prevents models from having enough states to capture the full history. However, we can limit the complexity in special cases. In particular, if we know that the probability of emitting a word x_t from a state z_t is 0, i.e. $p(x_t \mid z_t) = 0$ then we can ignore that state during inference.

Inspired by cloned HMMs (Dedieu et al., 2019), we add explicitly constrain the model to ensure that this properties holds at every time step. Specifically, we ensure that for an observed token x a fixed number of states z have $p(x \mid z) > 0$. For each word, we call this set $\mathcal{C}_x \subset \mathcal{Z}$. This yields the following constrained emission distribution:

$$p(x \mid z) \propto 1(z \in \mathcal{C}_x) e^{\phi_{xz}} \tag{3}$$

This allows us to perform marginal inference as follows

$$\begin{aligned} &p(\mathbf{x}) \\ &= \sum_{z_0} p(z_0) p(x_0 \mid z_0) \sum_{z_1} p(z_1 \mid z_0) p(x_1 \mid z_1) \\ &\quad \cdots \sum_{z_T} p(z_T \mid z_{T-1}) p(x_T \mid z_T) \\ &= \sum_{z_0 \in \mathcal{C}_{x_0}} p(z_0) p(x_0 \mid z_0) \sum_{z_1 \in \mathcal{C}_{x_1}} p(z_1 \mid z_0) p(x_1 \mid z_1) \\ &\quad \cdots \sum_{z_T \in \mathcal{C}_{x_T}} p(z_T \mid z_{T-1}) p(x_T \mid z_T) \end{aligned} \tag{4}$$

The key step in the second equality was only summing over the respective z_t with nonzero $p(x_t \mid z_t)$ for each x_t. This allows us to compute $p(\mathbf{x})$ with T matrix-vector products, each matrix of dimension $|\mathcal{C}_{x_t-1}| \times |\mathcal{C}_{x_t}|$. We choose sets $\mathcal{C}_x$ such that $\forall x, |\mathcal{C}_x| = k$ yielding a computation complexity of $O(Tk^2)$ rather than $O(T|\mathcal{Z}|^2)$. This gives a massive speedup in practice, for example if $|\mathcal{Z}|$ is on the order of 10k and k in the 100s. Please refer to Fig. 2 for an illustration of this method. (Describe here)

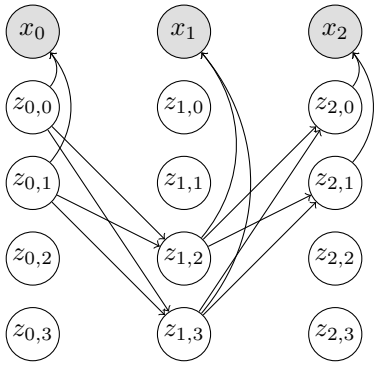

Figure 2: WILL REPLACE WITH TRELLIS A depiction of the emission sparsity constraints.

This method is a generalization of the one introduced in the cloned HMM (Dedieu et al., 2019), as we constrain the emission distribution such that every token is emit by k states, but a given state may emit more than one token. The cloned HMM constrains states to emit only a single token, which prevents its application to problems where the token space is larger than the state space.

4.2 State Dropout

We introduce a variant of dropout called state dropout that operates on the emission distribution. The goal of state dropout is to encourage usage of the full state space in HMMs as well as to reduce the complexity of marginalization in a manner that compounds with the sparse emission constraints discussed above.

State dropout prevents a token x from always being emit by a specific state z by zeroing out emission probabilities during training. Recall that the HMM has distributional parameters $\phi \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Z}|}$ for the emission matrix, and the emission sparsity constraints associate each token x with a set of states $\mathcal{C}_x$ that are the only states which emit x with nonzero probability. State dropout samples a mask over states $\mathbf{b}_x \in \{0,1\}^{|\mathcal{C}_x|}$ such that $|\mathbf{b}_x| = n$ for each x. Given the mask, we then computes the following emission distribution:

$$p(x \mid z) \propto b_{xz} 1(z \in \mathcal{C}_x) e^{\phi_{xz}} \tag{5}$$

Performing marginal inference with state dropout requires $O(Tn^2)$ computation, where n is the number of nonzero elements of $\mathbf{b}_{\mathcal{C}}, \forall \mathcal{C}$.

4.3 Parameterization

We parameterize the transition and emission distributions using a residual network with LayerNorm. Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^h$, where h is the hidden dimension of our neural network and the size of our token and state embeddings. We use the following neural network for $i \in \{\phi, \psi\}$

$$\begin{aligned} f_i(A) &= g(\text{ReLU}(W_{i1}A)) \\ g_i(B) &= \text{LayerNorm}(\text{ReLU}(W_{i2}B) + B) \end{aligned} \tag{6}$$

to define the distributional parameters, where $A, B \in \mathbb{R}^{h\times|\mathcal{Z}|}$ and $W_{i1}, W_{i2} \in \mathbb{R}^{h\times h}$. Let $V_x \in \mathbb{R}^{h\times|\mathcal{X}|}$ be the token embeddings and $U_{\text{previous}}, U_{\text{current}}, U_{\text{preterminal}} \in \mathbb{R}^{h\times|\mathcal{Z}|}$ the state embeddings. The distributional parameters are given by

$$\begin{aligned} \phi &= V_x^\top f_\phi(U_{\text{preterminal}}) \\ \psi &= U_{\text{current}}^\top f_\psi(U_{\text{previous}}) \end{aligned} \tag{7}$$

where $\phi \in \mathbb{R}^{|X|\times|Z|}$ and $\psi \in \mathbb{R}^{|Z|\times|Z|}$.

Compared to fully representing ϕ, ψ with a scalar parameterization which would require $O(|\mathcal{Z}|^2 + |\mathcal{X}||\mathcal{Z}|)$, this neural parameterization takes $O(h^2 + h|\mathcal{Z}| + h|\mathcal{X}|)$ parameters where h is in the 100s and $|\mathcal{Z}|, |\mathcal{X}|$ are on the order of 10k.

5 Experiments

In this section we provide further details on the application of the methods described in the previous section to language modeling and sequence tagging.

5.1 Language modeling

In word-level language modeling the tokens $\mathbf{x}$ correspond to the words $\mathbf{w} = \langle w_1, \ldots, w_T \rangle$ in a sentence. We thus learn a model over words and states $p(\mathbf{w}, \mathbf{z}) = \prod_t p(w_t \mid z_t) p(z_t \mid z_{t-1})$.

We use Brown clusters (Brown et al., 1992) to create the emission constraints. Brown clusters are obtained by assigning every token type in $\mathcal{X}$ a state in a HMM, then continually merging states until a desired number of states is reached. The merge at each iteration attempts to merge the two states that results in best likelihood.

To apply the Brown clusters as an emission constraint, each Brown cluster cluster is assigned a disjoint set of k states such that the state space is

partitioned. Each state corresponding to a Brown cluster is then allowed to emit only the words in that Brown cluster. This constraint corresponds to learning a refinement of the Brown clusters by splitting each cluster into k states.

Since the emission constraints using Brown clusters partitions the state space, we share state dropout masks for all word types within a Brown cluster.

5.2 Sequence Tagging

We extend the language modeling HMM to part of speech (POS) tagging, where the tokens $\mathbf{x}$ are now the product of words $\mathbf{w}$ as well as tags $\mathbf{y} = \langle y_1, \ldots, y_T \rangle$.

At every timestep, our model emits both a word w_t and tag y_t independently given the states z_t. This yields the following joint distribution

$$p(\mathbf{w}, \mathbf{y}, \mathbf{z}) = \prod_t p(w_t \mid z_t) p(y_t \mid z_t) p(z_t \mid z_{t-1}) \tag{8}$$

We use the same emission constraints and dropout strategy as language modeling, but applied only to the word emission distribution $p(w_t \mid z_t)$. The tag emission distribution is unconstrained.

Parameterization, CharCNN

5.3 Datasets

Language modeling We evaluate on the `Penn Treebank` (Marcus et al., 1993) and `wikitext2` (Merity et al., 2016) datasets. `Penn Treebank` contains 929k tokens in the training corpus, with a vocabulary size of 10k. We use the preprocessing from Mikolov et al. (2011), which lowercases all words and substitutes words outside of the vocabulary with unks. `Wikitext2` contains 2M tokens in the training corpus, with a vocabulary size of 33k. Casing is preserved, and all words outside the vocab are unked. Both datasets contain inter-sentence dependencies, due to the use of documents consisting of more than a single sentence.

Part of speech tagging We use the Wall Street Journal portion of the `Penn Treebank` for evaluating POS tagging. We map all numbers to 0, and perform no further preprocessing. We train on sections 0-18, validate on sections 19-21, and test on 22-24 following Ma and Hovy (2016).

Table 1: Perplexities on the `Penn Treebank` dataset. The FF model is a 256-dim 2-layer feedforward neural network with a window of 4 previous tokens with 0.3 dropout. The LSTM is a 256-dim 2-layer recurrent neural network with 0.3 dropout. The HMM is a 64k-state HMM with 0.5 state dropout.

Model	Num Params	Valid PPL	Test PPL
FF	2.8M	164	-
LSTM	3.6M	97	-
HMM	19.8M	122	-
HMM	7.7M	125	-

Table 2: Perplexities on the `wikitext2` dataset. The FF model is a 256-dim 2-layer feedforward neural network with a window of 4 previous tokens with 0.3 dropout. The LSTM is a 256-dim 2-layer recurrent neural network with 0.3 dropout. The HMM is a 32k-state HMM with 0.5 state dropout.

Model	Num params	Valid PPL	Test PPL
FF		209	-
LSTM		125	-
HMM		167	-

6 Results

We report perplexities for `Penn Treebank` in Table 1 and for `wikitext2` in Table 2.

The HMMs in all cases outperform the feedforward models (FF), but underperform the recurrent LSTMs. Since the HMMs require parameters linear in the number of hidden states, we find that the performance of the HMMs scales poorly compared to the other models which only require parameters that scale linearly with the vocbaulary size. Although representing and summing over the hidden states allows us to explicitly capture uncertainty in the hidden state, it proves to be a limiting factor in terms of performance.

In the remainder of this section we ablate and analyze the HMMs on `Penn Treebank`.

Sparse emission constraint ablation We ablate the emission sparsity constraints in Table 3, and find that the Brown emission constraints outperforms the uniform emission constraints in all model sizes.

One explanation for the relative benefit of Brown emission constraints over uniform is due to the effect of Brown clusters. The goal of Brown clus-

Table 3: The perplexities for the different emission sparsity constraints in a 1024 state HMM as well as larger HMMs for which exact inference without sparsity is too expensive. The quantities $|\mathcal{Z}|$ and k are the number of hidden states and the number of states per word respectively. The HMMs with 1024 states do not have any dropout, while the 8k and 16k state HMMs have unstructured dropout at a rate of 0.5.

Constraint	$\lvert\mathcal{Z}\rvert$	k	Valid PPL
Uniform	1k	32	-
Brown	1k	32	-
None	1k	1k	-
Uniform	8k	128	150*
Brown	8k	128	142*
Uniform	16k	128	146*
Brown	16k	128	134*

Table 4: The average entropies of the the emission and transition distributions for HMMs with uniform and Brown cluster emission constraints. All models have $k = 128$ states per word and use unstructured dropout with a rate of $p = 0.5$.

Constraint	$\lvert\mathcal{Z}\rvert$	H(emit)	H(transition)
Uniform	8k	-	-
Brown	8k	-	-
Uniform	16k	-	-
Brown	16k	-	-

tering is to place two words in the same cluster if they are used in the same context. In Table 4, we observe that the entropy of the emission distribution for models with uniform emission constraints is lower than models with brown constraints, and the entropy of the transition distributions is higher. This implies that the burden of modeling ambiguity is pushed onto the transition distribution, since the uniform models are less likely to place words that appear in similar contexts together.

Dropout analysis The effect of the different dropout strategies and rates is shown in Table 5. We found that state dropout outperformed unstructured dropout overall. This in addition to the computational benefits afforded by state dropout led us to prefer this strategy.

Unstructured dropout was sensitive to batching strategies. We observed a larger gap between training and validation perplexities with the unshuffled batching strategy, whereas unstructured dropout appeared to be robust to the batching strategy.

Table 5: State occupancies for the dropout strategies and rates. All models were HMMs with Brown cluster emission constraints, 16k total states, and 128 states per word (and therefore 128 Brown clusters).

Dropout type	p	Valid PPL	Occupancy
Unstructured	0.25	-	-
Unstructured	0.5	-	-
Unstructured	0.75	-	-
State	0.25	-	-
State	0.5	-	-
State	0.75	-	-

Table 6: The perplexities for HMMs with 16k states and different numbers of Brown clusters for constraining the emission distribution of the HMMs. $|\mathcal{Z}|$ is the total number of hidden states. All models have 0.5 state dropout.

$\lvert\mathcal{Z}\rvert$	Num clusters	Valid PPL
16k	32	136
16k	64	137
16k	128	133
16k	256	137

Analysis 1: Discuss what happens with just state dropout, without partitioning. (Too much variance in gradient estimator, would require variance reduction. Get numbers or just handwaive? Consider exact inference (in constrained model) this work.)

Number of Brown clusters We next examine the sensitivity of HMMs to the number of Brown clusters in Table 6. We find that models at with 16k or 32k total states are not sensitive to the number of Brown clusters in the range where marginalization is computationally feasible. This contrasts with the observation that the number of Brown clusters influenced performance at 1k total states.

Weight tying ablation Analysis 1: Discuss performance using different tying methods (found this to not affect performance) and computational savings in terms of number of parameters. Discuss relative parameter inefficiency compared to LSTM / FF.

State size ablation We find that as we increase the number of total states while keeping the Brown

Table 7: The perplexities for HMMs with 128 Brown clusters for constraining the emission distribution of the HMMs. $|\mathcal{Z}|$ is the total number of hidden states. All models have 0.5 state dropout.

$\|\mathcal{Z}\|$	Num clusters	Valid PPL
16k	128	133
32k	256	126
65k	512	124

clusters fixed, performance improves up until we have 65k states.

Analysis 1: How does state usage change as clusters remain constant but the number of states (and states per word) increases?

Parameterization ablation Analysis 1: Is there a performance difference between neural / scalar parameterization, and is it consistent across state sizes?

Discussion 1: Memory savings when working with state dropout to not instantiate the full matrices during training, which makes working with a neural parameterization beneficial.

6.1 Error analysis

7 Discussion

Acknowledgments

TBD

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155.

James Bradbury, Stephen Merity, Caiming Xiong, and Richard Socher. 2016. Quasi-recurrent neural networks. *CoRR*, abs/1611.01576.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4):467–479.

Jan Buys, Yonatan Bisk, and Yejin Choi. 2018. Bridging hmms and rnns through architectural transformations.

Antoine Dedieu, Nishad Gothoskar, Scott Swingle, Wolfgang Lehrach, Miguel Lázaro-Gredilla, and Dileep George. 2019. Learning higher-order sequential structure with cloned hmms.

Zoubin Ghahramani and Michael I. Jordan. 1997. Factorial hidden markov models. *Mach. Learn.*, 29(2–3):245–273.

Zhongqiang Huang. 2011. *Modeling Dependencies in Natural Languages with Latent Variables*. Ph.D. thesis, University of Maryland.

Yoon Kim, Chris Dyer, and Alexander M. Rush. 2019. Compound probabilistic context-free grammars for grammar induction. *CoRR*, abs/1906.10225.

Viktoriya Krakovna and Finale Doshi-Velez. 2016. Increasing the interpretability of recurrent neural networks using hidden markov models.

Thomas Kuhn, Heinrich Niemann, and Ernst Günter Schukat-Talamazzini. 1994. Ergodic hidden markov models and polygrams for language modeling. pages 357–360.

Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *CoRR*, abs/1603.01354.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and optimizing LSTM language models. *CoRR*, abs/1708.02182.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *CoRR*, abs/1609.07843.

Tomas Mikolov, Anoop Deoras, Stefan Kombrink, Lukás Burget, and Jan Cernocký. 2011. Empirical evaluation and combination of advanced language modeling techniques. pages 605–608.

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. pages 1045–1048.

Anjan Nepal and Alexander Yates. 2013. Factorial hidden markov models for learning representations of natural language. *CoRR*, abs/1312.6168.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. page 433–440.

Lawrence R. Rabiner. 1990. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, page 267–296. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

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

Ke M. Tran, Yonatan Bisk, Ashish Vaswani, Daniel Marcu, and Kevin Knight. 2016. Unsupervised neural hidden markov models. *CoRR*, abs/1609.09007.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, page 836–841, USA. Association for Computational Linguistics.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *CoRR*, abs/1409.2329.

A Hyperparameters

LSTM

- 2 layers

B Supplemental Material