Supplementary material for the paper Dynamic Neural Language Models

Edouard Delasalles, Sylvain Lamprier, and Ludovic Denoyer

Sorbonne Université, LIP6, F-75005, Paris, France
firstname.lastname@lip6.fr

A Full ELBO derivation

The generative model of DRLM writes as follow:

$$p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathcal{D},\mathbf{Z}) = \prod_{i=1}^{N} p_{\boldsymbol{\theta}}(\mathbf{d}^{(i)}|\mathbf{z}_{\mathrm{t}^{(i)}}) \prod_{t=0}^{T-1} p_{\boldsymbol{\psi}}(\mathbf{z}_{t+1}|\mathbf{z}_t).$$

We use use Variational Inference (VI) to learn this model, and consider a variational distribution $q_{\boldsymbol{\phi}}(\mathbf{Z})$ that factorizes across all timesteps:

$$q_{\boldsymbol{\phi}}(\mathbf{Z}) = \prod_{t=1}^{T} q_{\boldsymbol{\phi}}^{t}(\mathbf{z}_t),$$

where $q_{\boldsymbol{\phi}}^{t}$ are independent Gaussian distributions $\mathcal{N}(\mu_t, \sigma_t^2)$ with diagonal covariance matrices σ_t^2, and $\boldsymbol{\phi}$ is the total set of variational parameters.

A particularity of our approach is that we have several documents published at the same timestep. So, to obtain an Evidence Lower Bound (ELBO) we adapt the derivation in [3] as follows:

$$
\begin{aligned}
\log p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathcal{D}) &= \log \int_{\mathbf{Z}} p_{\boldsymbol{\psi}}(\mathbf{Z}) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\mathcal{D}_t|\mathbf{z}_t) d\mathbf{Z} \\
&= \log \int_{\mathbf{Z}} q_{\boldsymbol{\phi}}(\mathbf{Z}) p_{\boldsymbol{\psi}}(\mathbf{Z}) \frac{\prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\mathcal{D}_t|\mathbf{z}_t)}{q_{\boldsymbol{\phi}}(\mathbf{Z})} d\mathbf{Z} \\
&\geq \int_{\mathbf{Z}} q_{\boldsymbol{\phi}}(\mathbf{Z}) \log \left(p_{\boldsymbol{\psi}}(\mathbf{Z}) \frac{\prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\mathcal{D}_t|\mathbf{z}_t)}{q_{\boldsymbol{\phi}}(\mathbf{Z})} \right) d\mathbf{Z} \\
&= \sum_{t=1}^{T} \int_{\mathbf{z}_t} q^t_{\boldsymbol{\phi}}(\mathbf{z}_t) \log p_{\boldsymbol{\theta}}(\mathcal{D}_t|\mathbf{z}_t) d\mathbf{z}_t \\
&+ \sum_{t=1}^{T} \int_{\mathbf{z}_{t-1}} q^{t-1}_{\boldsymbol{\phi}}(\mathbf{z}_{t-1}) \int_{\mathbf{z}_t} q^t_{\boldsymbol{\phi}}(\mathbf{z}_t) \log \frac{p_{\boldsymbol{\psi}}(\mathbf{z}_t|\mathbf{z}_{t-1})}{q^t_{\boldsymbol{\phi}}(\mathbf{z}_t)} d\mathbf{z}_{t-1} d\mathbf{z}_t \\
&= \sum_{t=1}^{T} \mathbb{E}_{q^t_{\boldsymbol{\phi}}(\mathbf{z}_t)} \left[\log p_{\boldsymbol{\theta}}(\mathcal{D}_t|\mathbf{z}_t)\right] - \sum_{t=1}^{T} \mathbb{E}_{q^{t-1}_{\boldsymbol{\phi}}(\mathbf{z}_{t-1})} \left[D_{\mathrm{KL}}(q^t_{\boldsymbol{\phi}}(\mathbf{z}_t) \| p_{\boldsymbol{\psi}}(\mathbf{z}_t|\mathbf{z}_{t-1}))\right] \\
&= \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi}),
\end{aligned} \tag{1}
$$

where $\mathcal{D}_t$ is the set of all documents published at timestep t, and the inequality is obtained thanks to the Jensen theorem on concave functions.

The KL between two Gaussians owns an analytically closed form. This allows us to rewrite our log-likelihood lower-bound, noted $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi})$, as follows:

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi}) = &\sum_{t=1}^{T} \mathbb{E}_{q^t_{\boldsymbol{\phi}}(\mathbf{z}_t)} \left[p_{\boldsymbol{\theta}}(\mathcal{D}_t|\mathbf{z}_t)\right] + \frac{Td}{2} \\
&- \frac{1}{2} \Bigg(T \sum_{i=0}^{d_z-1} \log \sigma_i^2 - \sum_{t=1}^{T} \sum_{i=0}^{d_z-1} \log \eta_{t,i}^2 + \sum_{t=1}^{T} \sum_{i=0}^{d_z-1} \frac{\eta_{t,i}^2}{\sigma_i^2} \\
&+ \sum_{t=1}^{T} \mathbb{E}_{q^{t-1}_{\boldsymbol{\phi}}(\mathbf{z}_{t-1})} \left[(g(\mathbf{z}_{t-1}; \boldsymbol{w}) - \mu_t)'(\sigma^2)^{-1}(g(\mathbf{z}_{t-1}; \boldsymbol{w}) - \mu_t)\right] \Bigg)
\end{aligned}
$$

where we note A' the matrix transpose of a matrix A and where σ_i^2 and $\eta_{t,i}^2$ stand for the i-th component of diagonals σ^2 and η_t^2 respectively. This re-writing improve learning stability w.r.t. a version in which sampling would be done on the KL components too. Since the observation model p_θ is an RNN, the model is non-conjugate, and the ELBO in Equation 1 cannot be computed in closed form. We thus use the re-parametrization trick [2,4] to learn the model. It allows us to learn all parameters of the model jointly.

B Quantitative results for prediction

We report here the quantitative results for the prediction configuration on the language modeling task on 1.

Table 1: Prediction perplexity

Corpus	**S2**		**NYT**		**Reddit**	
Perplexity	*micro*	*macro*	*micro*	*macro*	*micro*	*macro*
LSTM	84.7	82.7	128.5	128.4	125.8	126.1
DT	92.0	89.6	137.1	137.0	151.1	151.6
DWE	87.0	84.8	140.1	140.0	136.5	139.9
DRLM-Id	81.2	79.2	123.7	123.6	124.7	125.0
DRLM	**79.7**	**77.8**	**123.3**	**123.1**	**123.9**	**124.3**

C Deriving Temporal Word Embedding Methods for Recurrent Language Modeling

We detail here how we adapt temporal word embeddings baselines to recurrent language modeling. The baselines are Dynamic Word Embeddings (DWE) [1], and DiffTime [5]. For both methods, we get rid of the context embeddings and only keep word embeddings $\mathbf{U}$.

C.1 Dynamic Word Embeddings

In DWE [1], Gaussian word embeddings are learned at each timestep with a temporal diffusion prior:

$$\mathbf{U}_{t+1}|\mathbf{U}_t \sim \mathcal{N}\left(\frac{U_t}{1+\sigma_t^2/\sigma_0^2}, \frac{1}{\sigma_t^{-2}+\sigma_0^{-2}}I\right),$$

where σ_0^2 and σ_t^2 are hyperparameters of the model.

We derive their skip-gram algorithm for our setting by maximizing the following approximate ELBO:

$$\begin{aligned}\mathcal{L}_{\mathcal{DWE}}(\boldsymbol{\theta},\boldsymbol{\phi}) = \sum_{t=1}^{T} & \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{U}_t)}\left[\log p_{\boldsymbol{\theta}}(\mathbf{X}^t|\mathbf{U}^t)\right] + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{U}_t)}\left[\log \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{U}_{t-1})}\left[p(\mathbf{U}_t|\mathbf{U}_{t-1})\right]\right] \\ & - \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{U}_t)}\left[\log q_{\boldsymbol{\phi}}(\mathbf{U}_t)\right],\end{aligned} \tag{2}$$

where $p_{\boldsymbol{\theta}}$ is paramatrized by an LSTM. $q_{\boldsymbol{\phi}}$ is a variational Gaussian distribution that factorizes as:

$$q_{\boldsymbol{\phi}}(\mathbf{U}) = \prod_{t=1}^{T} q_{\boldsymbol{\phi}}(\mathbf{U}_t),$$

and $\boldsymbol{\phi}$ are its parameters.

To learn this model, we sample a mini-batches $\mathbf{M}$ that contains text coming from different training timesteps. We must hence rescale the ELBO in Equation 2. We do so by estimating the probability that a given word appears in a particular mini-batch:

$$\begin{aligned}\mathcal{L}_{minibatch}(\boldsymbol{\theta}, \boldsymbol{\phi}) = &\frac{|\mathbf{X}|}{|\mathbf{M}|} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{U}^{\mathbf{M}})} \left[\sum_{\mathbf{x} \in \mathbf{M}} \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{U}^{\mathbf{M}}) \right] \\ &+ \sum_{\mathbf{u} \in \mathbf{U}^{\mathbf{M}}} \frac{1}{(1-(1-\nu_{\mathbf{u}})^{|\mathbf{M}|})} \sum_{t=1}^{T} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{u})} \left[\log \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{u}_{t-1})} \left[p(\mathbf{u}_t|\mathbf{u}_{t-1}) \right] \right] \\ &- \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{u}_t)} \left[\log q_{\boldsymbol{\phi}}(\mathbf{u}_t) \right],\end{aligned}$$

where $\mathbf{U}^{\mathbf{M}}$ are the embeddings of words in $\mathbf{M}$, $\nu_{\mathbf{u}}$ is the apparition frequency of term whose embedding is $\mathbf{u}$ in $\mathbf{X}$, and $|\mathbf{X}|$ (respectively $|\mathbf{M}|$) is the number of words in $\mathbf{X}$ ($\mathbf{M}$). In this formulation, gradient computation does not require any approximation, while allowing it to flow through all timesteps.

C.2 DiffTime

The adaptation of the DiffTime baseline [5] is straightforward. It learns a non-linear function d that outputs temporal word embeddings:

$$\mathbf{u}_t = d(\mathbf{u}, t; \boldsymbol{\phi})$$

where $\mathbf{u}$ is a learned word embedding, t is a scalar timestep, and $\boldsymbol{\phi}$ are the function's parameters. We refer the reader to the complete paper for more details on the implementation of d.

For recurrent language modeling adaptation, we simply learn jointly the word embeddings $\mathbf{U}$, the parameters $\boldsymbol{\phi}$ of d and the parameters $\boldsymbol{\theta}$ of an LSTM by maximizing the following likelihood:

$$\mathcal{L}_{\mathcal{DT}}(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{U}) = \prod_{t=1}^{T} \prod_{\mathbf{x} \in \mathbf{X}^t} \prod_{k=1}^{|\mathbf{x}|-1} p_{\boldsymbol{\theta}}(\mathrm{x}_{k+1}|\mathbf{u}^t_{1:k}).$$

References

1. Bamler, R., Mandt, S.: Dynamic word embeddings. In: ICML (2017)
2. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)

3. Krishnan, R.G., Shalit, U., Sontag, D.: Structured inference networks for nonlinear state space models. In: AAAI (2017)
4. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: ICML (2014)
5. Rosenfeld, A., Erk, K.: Deep neural models of semantic shift. In: NAACL (2018)