1 Summary

This is a note for [1]. The implementation code for this paper can be found at 1.

The important components are as follows:

1. RNN language model
2. Morphological priors
3. Latent word embedding $b_w$.
4. Morpheme embedding $u_m$.
5. Variational distribution $Q(b)$

2 Latent Word Embedding and Morpheme Embedding

Each morpheme is segmented in unsupervised fashion according to Morfessor. For example, $u_{-ism} = (-0.24, 5, -111)$.

When inferring $P(x)$, we will have to infer $P(b)$ too since $P(b)$ appears in the lower variational bound.

\[ b_{w,i} \sim Bernoulli(sigmoid(\sum_{m \in M_w} u_{m,i})) \]

i.e. for outcomes or the range of a probabilistic variable $b_{w,i}$ is either 0 or 1,

\[ P(b_{w,i}) = sigmoid(\sum_{m \in M_w} u_{m,i})^{b_{w,i}}(1 - sigmoid(\sum_{m \in M_w} u_{m,i}))^{1-b_{w,i}} \]

So let’s look into an example. Let $M = perfection$, $-ism$ $u_{perfection} = (0, -1, 1, 1)$

When $w = perfectionism$, then

$u_{-ism} = (2, 5, 1, 3)$

$b_{w,0} \sim Bernoulli(sigmoid(0 + 2)) \approx 0.88$

$b_{w,1} \sim Bernoulli(sigmoid(-1.1 + 5.1)) \approx 0.98$

$b_{w,2} \sim Bernoulli(sigmoid(1 + 3)) \approx 0.98$

So $P(b_w = (1, 1, 1)) = 0.88 \times 0.98 \times 0.98 \approx 0.84$.

3 Hidden state

The hidden state at time $h_t$ (vector) is

\[ h_t = sigmoid(\Theta h_{t-1} + b_{x_t}) \]

where $x_t$ is the word corresponding to the position $t$, and $\Theta$ is the parameter for the recurrence function (recurrent weights2).

1https://github.com/rguthrie3/MorphologicalPriorsForWordEmbeddings
2http://peterroelants.github.io/posts/rnn_implementation_part01/
4 What is going on inside $D_{KL}(Q(b) || P(b))$?

$$D_{KL}(q(b_{w,i})||P(b_{w,i})) = q(b_{w,i}) \log\left(\frac{q(b_{w,i})}{P(b_{w,i})}\right)$$

$$= q(b_{w,i}) \log(q(b_{w,i})) - \log(P(b_{w,i}))$$

$$= E_q[\log(q(b_{w,i}))] - E_q[\log(P(b_{w,i}))]$$

$$E_q[\log(q(b_{w,i};\gamma_{w,i}))] = q(b_{w,i} = 1) \log(\gamma_{w,i}) + q(b_{w,i} = 0) \log(1 - \gamma_{w,i})$$

$$= \gamma_{w,i} \log(\gamma_{w,i}) + (1 - \gamma_{w,i}) \log(1 - \gamma_{w,i})$$

$$E_q[\log P(b_{w,i})] = q(b_{w,i} = 1) \log(\text{sigmoid}(\sum_{m \in M_w} u_{m,i})) + q(b_{w,i} = 0) \log((1 - \text{sigmoid}(\sum_{m \in M_w} u_{m,i})))$$

$$= \gamma_{w,i} \log(\text{sigmoid}(\sum_{m \in M_w} u_{m,i})) + (1 - \gamma_{w,i}) \log((1 - \text{sigmoid}(\sum_{m \in M_w} u_{m,i})))$$

Note that `morpho_level_reps = (self.morpho_embed_lookup.apply(morpho_idxs) * masks).sum(axis=2)` represents $\sum_{m \in M_w} u_{m,i}$.

$$1 - \text{sigmoid}(x) = \frac{1 + e^{-x}}{1 + e^{-x}} - \frac{1}{1 + e^{-x}} = \frac{e^{-x}}{1 + e^{-x}} = \frac{1}{e^x + 1}$$

$$\log(1 - \text{sigmoid}(x)) = \log\left(\frac{1}{e^x + 1}\right) = -\log(e^x + 1)$$

References