What are effective ways of incorporating syntactic structure into neural language models?

		OPEN
		OPEN
$\vert$		GEN *The*
$\vert\vert$ *The*	*The*	GEN *hungry*
$\vert\vert$ *The* $\vert$ *hungry*	*The* $\vert$ *hungry*	GEN *cat*
$\vert\vert$ *The* $\vert$ *hungry* $\vert$ *cat*	*The* $\vert$ *hungry* $\vert$ *cat*	REDUCE
$\vert$ *The hungry cat*	*The* $\vert$ *hungry* $\vert$ *cat*	OPEN
$\vert$ *The hungry cat* $\vert$	*The* $\vert$ *hungry* $\vert$ *cat*	GEN *meows*
$\vert$ *The hungry cat* $\vert\vert$ *meows*	*The* $\vert$ *hungry* $\vert$ *cat* $\vert$ *meows*	REDUCE
$\vert$ *The hungry cat* $\vert$ *meows*	*The* $\vert$ *hungry* $\vert$ *cat* $\vert$ *meows*	GEN.
$\vert$ *The hungry cat* $\vert$ *meows* $\vert$	*The* $\vert$ *hungry* $\vert$ *cat* $\vert$ *meows* $\vert$	REDUCE
The hungry cat meows	*The* $\vert$ *hungry* $\vert$ *cat* $\vert$ *meows* $\vert$	

$a = \langle a_1, \dots, a_T \rangle yxx$

$p(a) = p(y \mid x) p(a) = p(x, y)$

$_D = \{\text{SHIFT}, \text{OPEN}, \text{REDUCE}\},_G = \{\text{GEN}, \text{OPEN}, \text{REDUCE}\}.\Lambda\mathcal{X}a\mathcal{A}_D a\mathcal{A}_G ayn = \langle n_1, \dots, n_K \rangle \Lambda^K yx\mathcal{X}^N \mu_a : \{1, \dots, T\} \to$

$a\mathcal{A}_D Ty \mid x) = p(a \mid x) = \prod_{t=1}^{T} P(a_t \mid x, a_{<t}),$

$n\mu(t)$

$a\mathcal{A}_G T^{14} y \mid x) = p(a) = \prod_{t=1}^{T} P(a_t \mid a_{<t}),$

$x\nu(t)$

$t\mathbf{u}_t{}^{1516}{}_t \mid a_{<t}) \propto \exp\left\{[\text{FFN}_\alpha(\mathbf{u}_t)]_{a_t}\right\}$

$p(n_{\mu(t)} \mid a_{<t}) \propto \exp\left\{[\text{FFN}_\beta(\mathbf{u}_t)]_{n_{\mu(t)}}\right\}$

$p(x_{\nu(t)} \mid a_{<t}) \propto \exp\left\{[\text{FFN}_\gamma(\mathbf{u}_t)]_{x_{\nu(t)}}\right\}\alpha\beta\gamma\mathbf{u}_t$

$_D = \{\text{REDUCE}, \text{SHIFT}\} \cup \{\text{OPEN}(n) \mid n \in \Lambda\},_G = \{\text{REDUCE}\} \cup \{\text{OPEN}(n) \mid n \in \Lambda\} \cup \{\text{GEN}(x) \mid x \in \mathcal{X}\},p(a)\mathcal{X}^{17}$

$\mathbf{u}_t t\mathbf{u}_t = [\mathbf{s}_t; \mathbf{b}_t; \mathbf{h}_t].\mathbf{s}_t\mathbf{b}_t\mathbf{h}_t$

$\mathbf{b}_t\mathbf{h}_t\mathbf{s}_t$

REDUCE

REDUCE

very hungry

$[_1, \dots, _m]m\mathbf{n}a_i i_i = \frac{a_i}{\sum_{i=1}^{m} \tilde{a}_i}$

$\tilde{a}_i = \exp\{_i^\top \mathbf{V}[\mathbf{u}_t; \mathbf{n}]\}_i \mathbf{u}_t \mathbf{n} \mathbf{V} \sum_{i=1}^{m} a_i \mathbf{h}_i \circ \mathbf{n} + (1 - \mathbf{g}) \circ \mathbf{m}.n^{18} \mathbf{g}[\mathbf{n}; \mathbf{m}]^{19}$

$\theta) = \sum_{(x,y)\in\mathcal{D}} \log p_\theta(y \mid x),\theta) = \sum_{(x,y)\in\mathcal{D}} \log p_\theta(x, y),\theta$

x *a posteriori* $\hat{y} =_{y\in\mathcal{Y}(x)} p_\theta(y \mid x).\hat{a}_t =_a p_\theta(a \mid \hat{a}_{<t}).y^* = yield(\hat{a})\hat{a} = \langle \hat{a}_1, \dots, \hat{a}_m \rangle$

$p_\theta x\hat{y} =_{y\in\mathcal{Y}(x)} p_\theta(x, y),x\sum_{y\in\mathcal{Y}(x)} p_\theta(x, y).q_\lambda(y \mid x)$

$x\sum_{y\in\mathcal{Y}(x)} p_\theta(x, y)$

$= \sum_{y\in\mathcal{Y}(x)} q_\lambda(y \mid x) \frac{p_\theta(x,y)}{q_\lambda(y|x)}$

$=_q \left[\frac{p_\theta(x,y)}{q_\lambda(y|x)}\right]$

$_q\left[\frac{p_\theta(x,y)}{q_\lambda(y|x)}\right] \approx \frac{1}{K}\sum_{i=1}^{K} \frac{p_\theta(x,y^{(i)})}{q_\lambda(y^{(i)}|x)},y^{(i)} q_\lambda x$

$\hat{y}yp_\theta(y, x)$

$xy \in \mathcal{Y}(x)xy \Rightarrow q(y \mid x) > 0.$

[20]

[21]

$88.47 \pm 0.17 (88.58)$

$91.07 \pm 0.1 (91.12) 91.02 \pm 0.05 (91.04) 93.32 \pm 0.1 (93.32)$

$108.76 \pm 1.52 (107.43) 107.80 \pm 1.59 (106.45)$

??

$H(Y \mid X = x) H(Y \mid X) p_X(x) p_{Y|X}$

[22]

[23]

n^{24}

$xy\mathcal{Y}(x)\Psi\Psi(x, y)y \mid x) = \frac{\Psi(x,y)}{Z(x)},x\sum_{y\in\mathcal{Y}(x)} \Psi(x, y)x$

$\Psi yy_a = (A, i, j)A\langle x_{i+1}, \dots, x_j \rangle y = \{y_a\}_{a=1}^{A} \Psi(x, y)\Psi(x, y) = \prod_{a=1}^{A} \psi(x, y_a),\psi(x, y_a)\psi y_a$

anchored rules *labeled spans*[25] ψ

$\psi y_a(A, i, j)y\psi(x, y_a) > 0x\mathbf{f}_i\mathbf{b}_i i(i, j)\mathbf{s}_{ij} = [\mathbf{f}_j - \mathbf{f}_i; \mathbf{b}_i - \mathbf{b}_j].\mathbf{s}_{ij} x_i^j R^\Lambda A \log\psi(x, y_a) = [\text{FFN}(\mathbf{s}_{ij})]_A,A$

$(1, 4)$ RNN