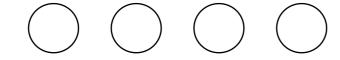
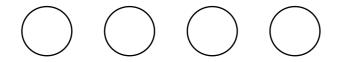
LM



1, 2, 3, 4

PLM



3, 2, 1, 4

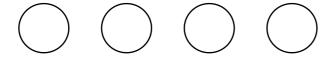
2, 3, 4, 1

LM





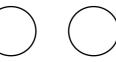
PLM



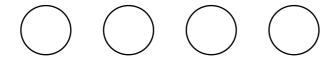
LM







PLM



LM









3, 2, 1, 4

PLM







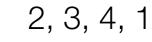




















LM









1, 2, 3, 4

PLM















3, 2, 1, 4

LM









1, 2, 3, 4

PLM









3, 2, 1, 4









2, 3, 4, 1









LM







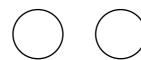
1, 2, 3, 4

PLM

















LM

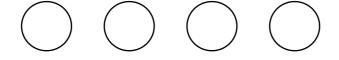


1, 2, 3, 4

PLM



3, 2, 1, 4



2, 3, 4, 1

LM

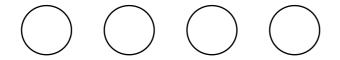


1, 2, 3, 4

PLM



3, 2, 1, 4



2, 3, 4, 1

LM



1, 2, 3, 4

PLM



3, 2, 1, 4



2, 3, 4, 1

LM



1, 2, 3, 4

PLM



3, 2, 1, 4



2, 3, 4, 1

LM



1, 2, 3, 4

PLM



3, 2, 1, 4



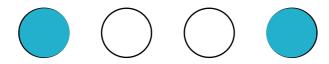
2, 3, 4, 1

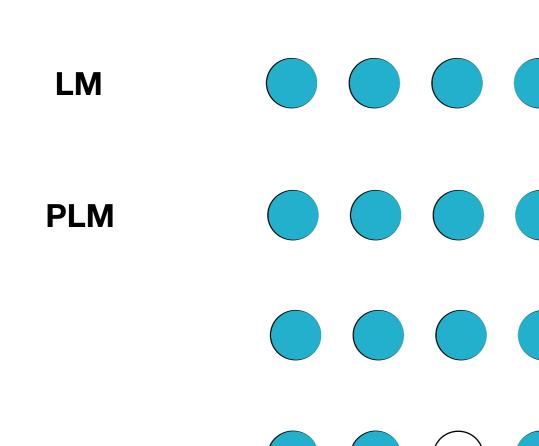
3, 2, 1, 4

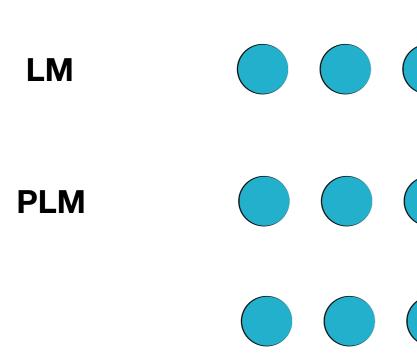
2, 3, 4, 1

LM **PLM**









3, 2, 1, 4

LM



1, 2, 3, 4

PLM

3, 2, 1, 4

2, 3, 4, 1

$$p_{ heta}(\mathbf{x}) = \max_{ heta} \quad \mathbb{E}_{\mathbf{z} \sim \mathcal{Z}_T} \left[\sum_{t=1}^T \log p_{ heta}(x_{z_t} \mid \mathbf{x}_{\mathbf{z}_{< t}}) \right]$$

LM



1, 2, 3, 4

PLM

3, 2, 1, 4

2, 3, 4, 1

$$p_{\theta}(\mathbf{x}) = \max_{\theta} \quad \mathbb{E}_{\mathbf{z} \sim \mathcal{Z}_T} \left[\sum_{t=1}^{T} \log p_{\theta}(x_{z_t} \mid \mathbf{x}_{\mathbf{z}_{< t}}) \right]$$

$$p_{\theta}(X_{z_t} = x \mid \mathbf{x}_{z_{< t}}) = \frac{\exp\left(e(x)^{\top} g_{\theta}(\mathbf{x}_{\mathbf{z}_{< t}}, z_t)\right)}{\sum_{x'} \exp\left(e(x')^{\top} g_{\theta}(\mathbf{x}_{\mathbf{z}_{< t}}, z_t)\right)}$$

XLNet Two-Stream Self-Attention

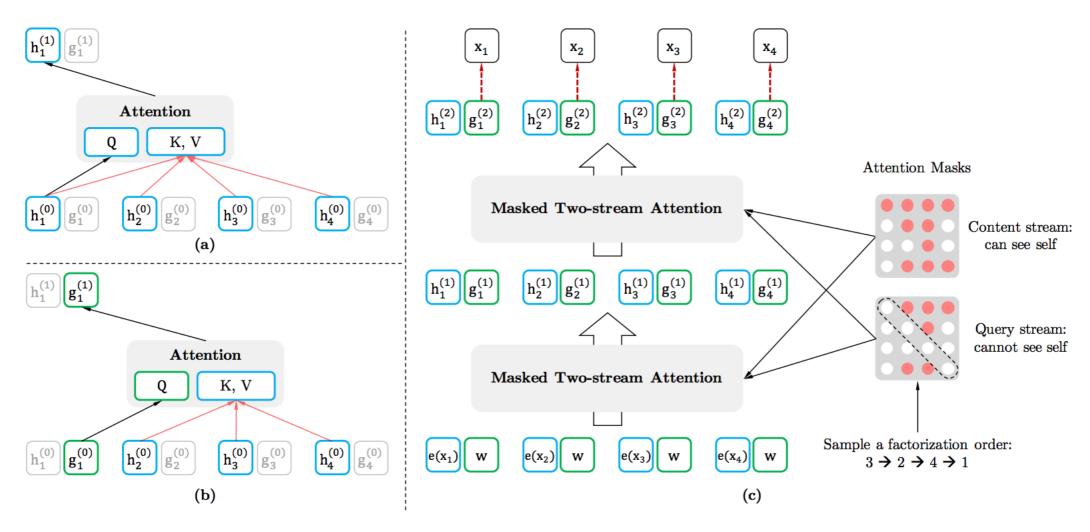
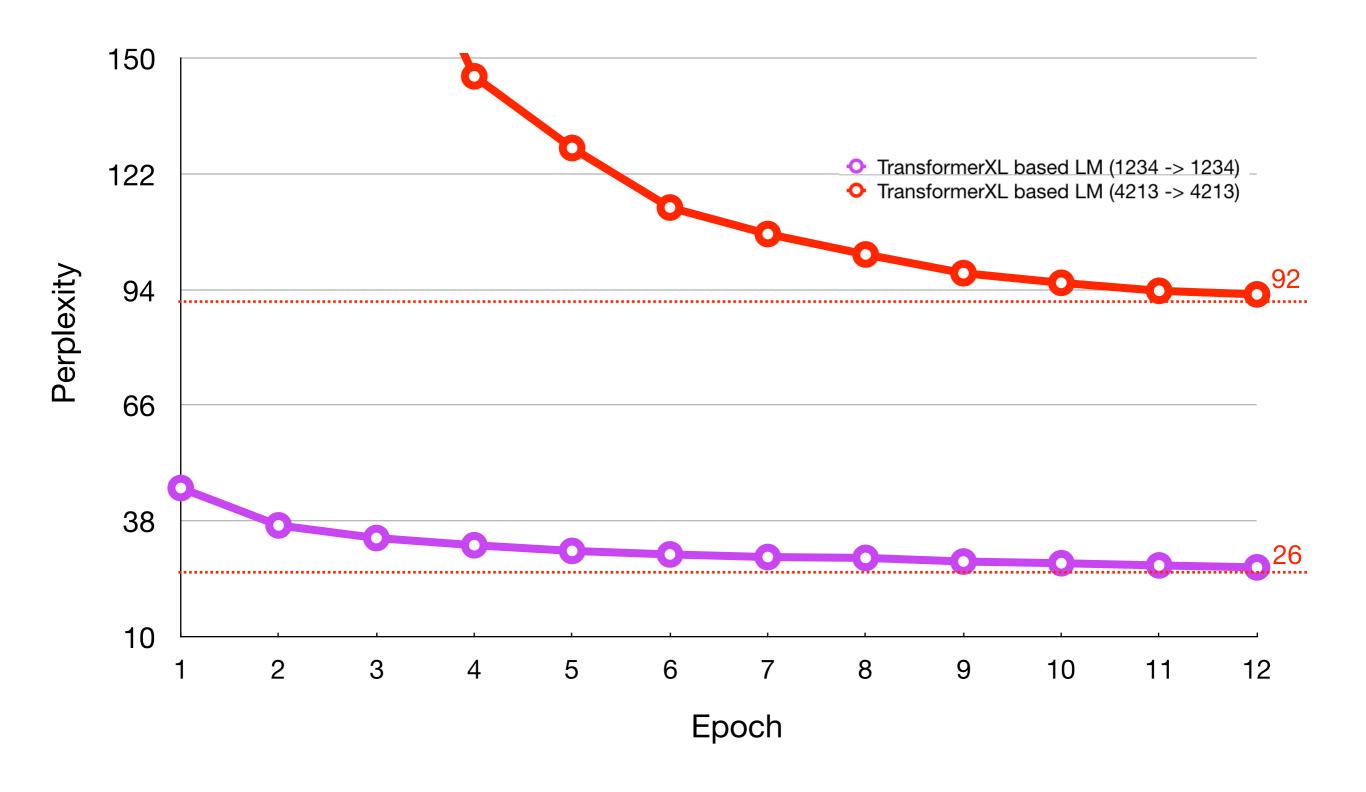


Figure 2: (a): Content stream attention, which is the same as the standard self-attention. (b): Query stream attention, which does not have access information about the content x_{z_t} . (c): Overview of the permutation language modeling training with two-stream attention.

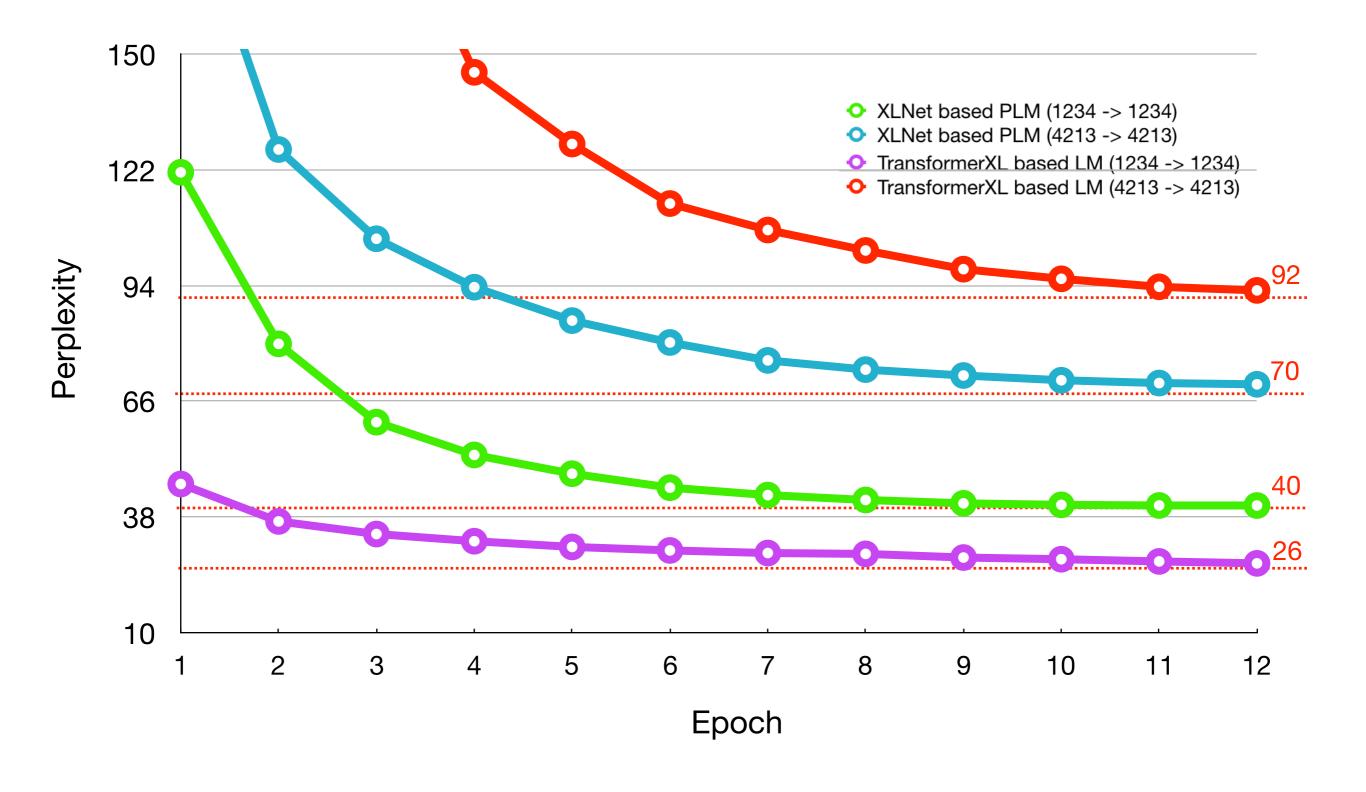
$$p_{\theta}(X_{z_{t}} = x \mid \mathbf{x}_{z_{< t}}) = \frac{\exp\left(e(x)^{\top} g_{\theta}(\mathbf{x}_{\mathbf{z}_{< t}}, z_{t})\right)}{\sum_{x'} \exp\left(e(x')^{\top} g_{\theta}(\mathbf{x}_{\mathbf{z}_{< t}}, z_{t})\right)} \qquad g_{z_{t}}^{(m)} \leftarrow \operatorname{Attention}(Q = g_{z_{t}}^{(m-1)}, KV = \mathbf{h}_{\mathbf{z}_{< t}}^{(m-1)}; \theta) \\ h_{z_{t}}^{(m)} \leftarrow \operatorname{Attention}(Q = h_{z_{t}}^{(m-1)}, KV = \mathbf{h}_{\mathbf{z}_{< t}}^{(m-1)}; \theta)$$

Yang, Zhilin, et al. "XLNet: Generalized Autoregressive Pretraining for Language Understanding." arXiv preprint arXiv:1906.08237 (2019).

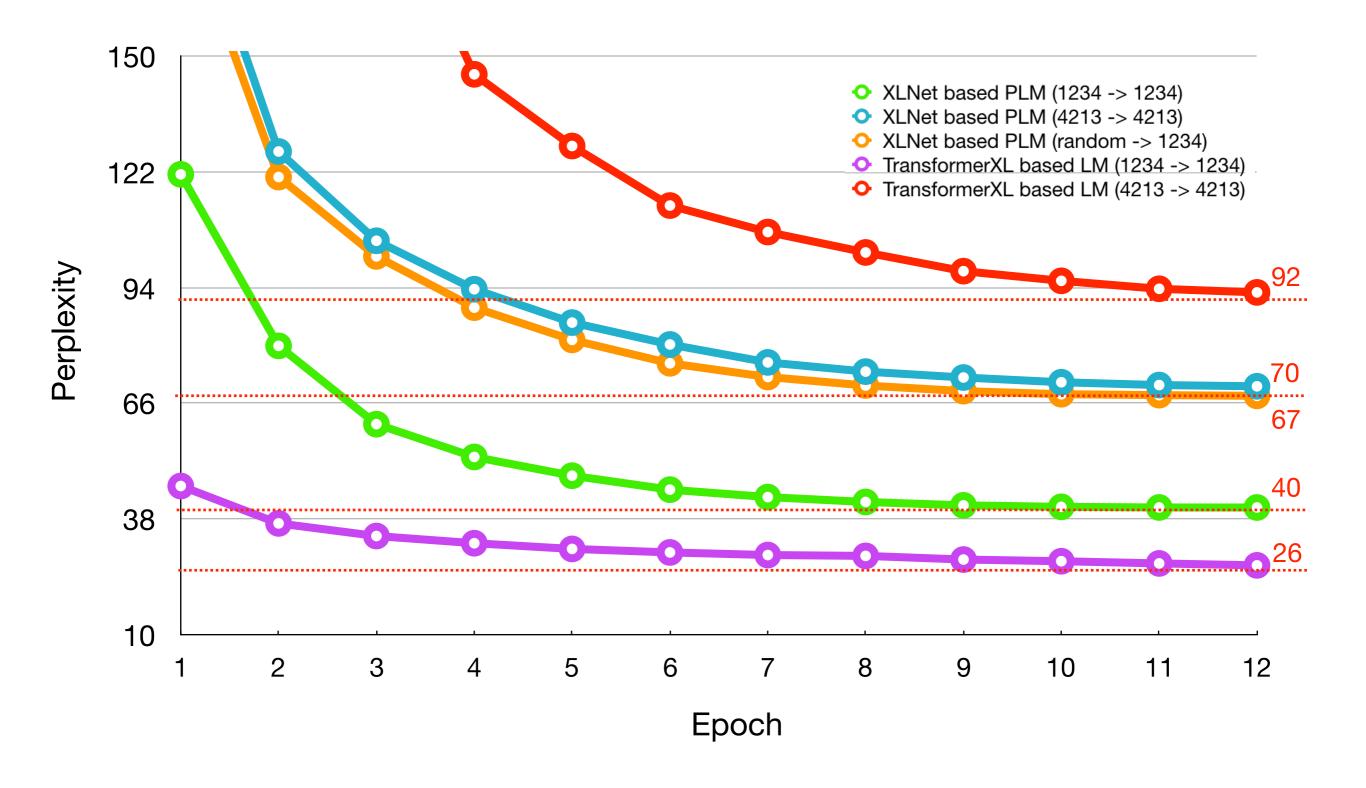
LM on Wiki-103: Sequential vs. Arbitrary



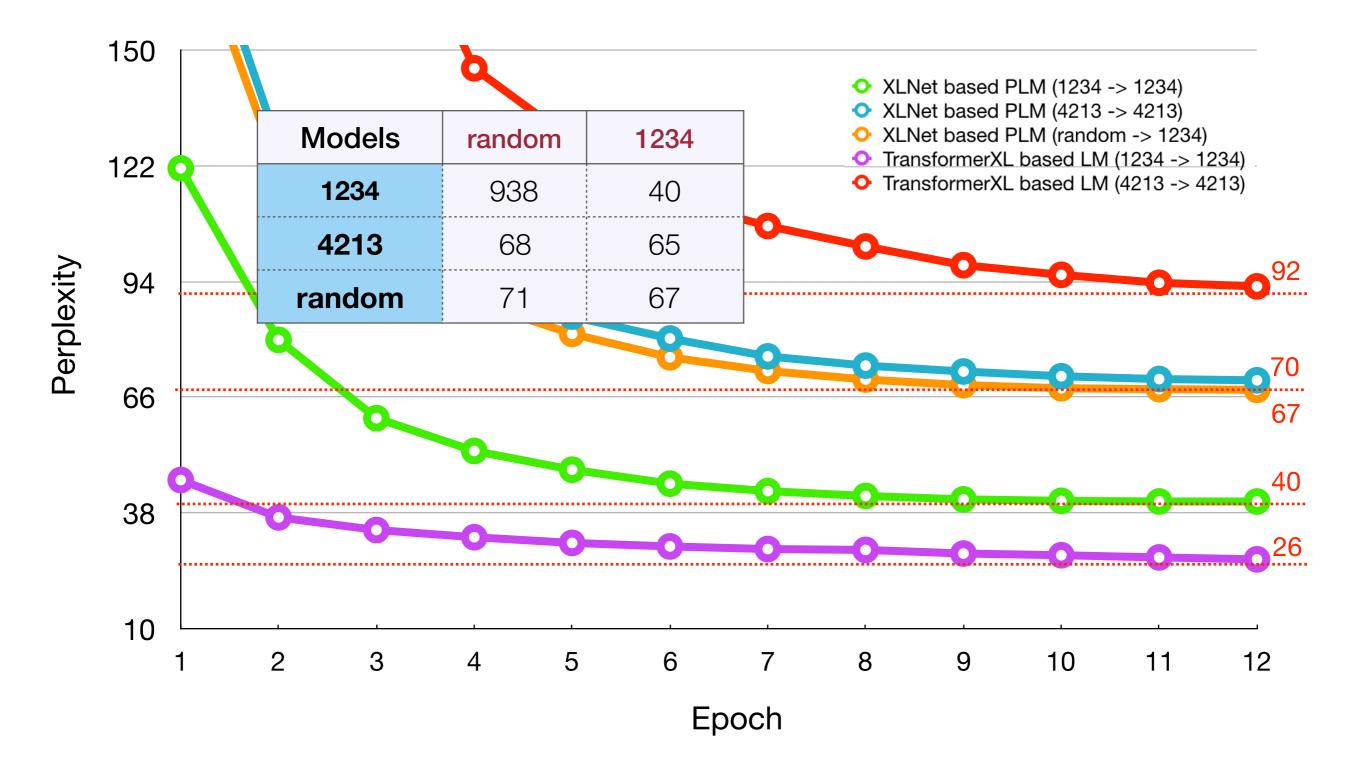
PLM on Wiki-103: Sequential vs. Arbitrary



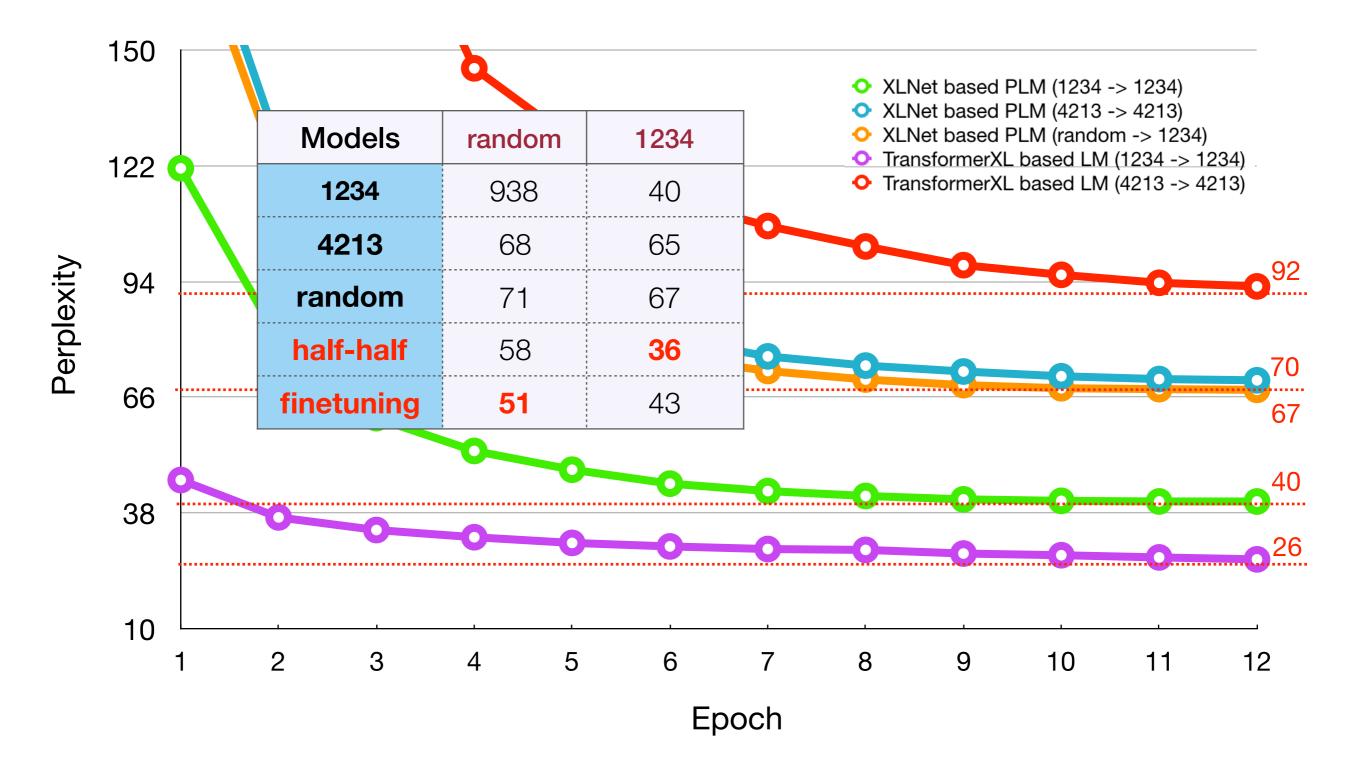
PLM on Wiki-103: Random Training



PLM Evaluation on Wiki-103



PLM Evaluation on Wiki-103



Does PLM help ASR Rescore?

LM Models	615	chat-0	chat-1	aishell	aishell_dev	aishell_mic	huoguo	spt
LSTM	1.555	13.622	20.366	2.938	4.028	4.090	19.607	17.576
ONLSTM	1.574	13.692	20.397	2.923	4.100	4.125	19.621	17.407
XLNet								

Note: time is short, this part of experiment is not finished.