A APPENDIX

A.1 A PROOF OF MLM ON CODE-SWITCHED DATASET

Proposition Training model with standard masked language modeling on source-language,target-language and source-target code-switched sentences is approximately optimizing:

$$\boldsymbol{c}_{x_i}^{src} \sim \boldsymbol{e}_{x_i} \sim \boldsymbol{e}_{y_i} \sim \boldsymbol{c}_{y_i}^{tgt}$$

where $\boldsymbol{c}_{x_i}^{src}$ represents the contextualized embedding of token x_i in source sentence and $\boldsymbol{c}_{y_i}^{src}$ represents the contextualized embedding of token y_i in target sentence. And $\boldsymbol{e}_{x_i}$ and $\boldsymbol{e}_{y_i}$ are word embeddings of token x_i and y_i in vocabulary.

Note Under the existing conditions, we can not derive a strict bound but an approximate conclusion. Given $\boldsymbol{a} \in \mathbb{R}^n, \boldsymbol{b} \in \mathbb{R}^n, \boldsymbol{c} \in \mathbb{R}^n$ we assume $\boldsymbol{a} \sim \boldsymbol{b}$, if the projection components of $\boldsymbol{a}$ and $\boldsymbol{b}$ onto another vector $\boldsymbol{c}$ are the same: $\boldsymbol{a} \cdot \boldsymbol{c} = \boldsymbol{b} \cdot \boldsymbol{c}$. We think this is approximately reasonable for word alignment task, because for word alignments method two words are aligned as long as their similarity is higher than other words in two parallel sentences and doesn't need to exceed a fixed number. And in section 3.3, we give a corollary. Subsequent experiments prove that our assumption is reasonable.

Proof We denote the embeddings of the corresponding original tokens as $\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_L$. The MLM objective $\mathcal{L}_{\mathrm{MLM}}(\boldsymbol{x})$ can be formulated as:

$$-\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \log \frac{\exp\left(\boldsymbol{m}_i \cdot \boldsymbol{e}_i\right)}{\sum_{k=1}^{|\mathcal{V}|} \exp\left(\boldsymbol{m}_i \cdot \boldsymbol{e}_k\right)} = -\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \log \sum_{k=1}^{|\mathcal{V}|} \exp\left(\boldsymbol{m}_i \cdot \boldsymbol{e}_k - \boldsymbol{m}_i \cdot \boldsymbol{e}_i\right) \tag{1}$$

where $\mathcal{M}$ denotes the set of masked tokens and $|\mathcal{V}|$ is the size of vocabulary $\mathcal{V}$. $\boldsymbol{m}_i$ is hidden state of the last layer at the masked position, and can be regarded as a fusion of contextualized representations of surrounding tokens. Given two sentences: one source-language sentence $\mathbf{x} = \langle x_1, \cdots, x_{i-1}, x_i, x_{i+1}, \cdots, x_n \rangle$ of length n and its code-switched sentence $\mathbf{x}' = \langle x_1, \cdots, x_{i-1}, y_i, x_{i+1}, \cdots, x_n \rangle$, where $\langle x_i, y_i \rangle$ is aligned pair. If we only mask x_i in the $\mathbf{x}$ and y_i in the $\mathbf{x}'$, then $\mathbf{x_{mask}} = \langle x_1, \cdots, x_{i-1}, < mask >, x_{i+1}, \cdots, x_n \rangle = \langle x_1, \cdots, x_{i-1}, < mask >, x_{i+1}, \cdots, x_n \rangle = \mathbf{x'_{mask}}$, the loss function can be written as

$$\mathcal{L}_{\mathrm{MLM}} = \mathcal{L}_{\mathbf{x}} + \mathcal{L}_{\mathbf{x}'} = -\frac{1}{2}(\log \sum_{k=1}^{|\mathcal{V}|} \exp\left(\boldsymbol{m} \cdot \boldsymbol{e}_k - \boldsymbol{m} \cdot \boldsymbol{e}_{x_i}\right) + \log \sum_{k=1}^{|\mathcal{V}|} \exp\left(\boldsymbol{m} \cdot \boldsymbol{e}_k - \boldsymbol{m} \cdot \boldsymbol{e}_{y_i}\right)) \tag{2}$$

This inequality below is easily proved.

$$\max\left\{x_1, \ldots, x_n\right\} \leq \log \sum_{i=0}^{n} e^{x_i} \leq \max\left\{x_1, \ldots, x_n\right\} + \log n \tag{3}$$

So for

$$\mathcal{L}_{\mathbf{x}} = -\log \sum_{k=1}^{|\mathcal{V}|} \exp\left(\boldsymbol{m} \cdot \boldsymbol{e}_k - \boldsymbol{m} \cdot \boldsymbol{e}_{x_i}\right) \tag{4}$$

We have:

$$\max \begin{pmatrix} \boldsymbol{m} \cdot \boldsymbol{e}_0 - \boldsymbol{m} \cdot \boldsymbol{e}_{x_i} \\ \vdots \\ \boldsymbol{m} \cdot \boldsymbol{e}_{x_i-1} - \boldsymbol{m} \cdot \boldsymbol{e}_{x_i} \\ 0 \\ \boldsymbol{m} \cdot \boldsymbol{e}_{x_i+1} - \boldsymbol{m} \cdot \boldsymbol{e}_{x_i} \\ \vdots \\ \boldsymbol{m} \cdot \boldsymbol{e}_{|\mathcal{V}|} - \boldsymbol{m} \cdot \boldsymbol{e}_{x_i} \end{pmatrix} \leq \log \sum_{k=1}^{|\mathcal{V}|} e^{\boldsymbol{m} \cdot \boldsymbol{e}_k - \boldsymbol{m} \cdot \boldsymbol{e}_{x_i}} \leq \max \begin{pmatrix} \boldsymbol{m} \cdot \boldsymbol{e}_0 - \boldsymbol{m} \cdot \boldsymbol{e}_{x_i} \\ \vdots \\ \boldsymbol{m} \cdot \boldsymbol{e}_{x_i-1} - \boldsymbol{m} \cdot \boldsymbol{e}_{x_i} \\ 0 \\ \boldsymbol{m} \cdot \boldsymbol{e}_{x_i+1} - \boldsymbol{m} \cdot \boldsymbol{e}_{x_i} \\ \vdots \\ \boldsymbol{m} \cdot \boldsymbol{e}_{|\mathcal{V}|} - \boldsymbol{m} \cdot \boldsymbol{e}_{x_i} \end{pmatrix} + \log n \tag{5}$$

In Ineq.5, 0 is fixed value. So when training model with this loss function, model is optimized to learn $\boldsymbol{m} \cdot \boldsymbol{e}_k - \boldsymbol{m} \cdot \boldsymbol{e}_{x_i} \leq 0, \forall k \in |\mathcal{V}|$. In other words, $\boldsymbol{m} \cdot \boldsymbol{e}_k \leq \boldsymbol{m} \cdot \boldsymbol{e}_{x_i}, \forall k \in |\mathcal{V}|$. When $k = y_i$,

we have $\boldsymbol{m} \cdot \boldsymbol{e}_{y_i} \leq \boldsymbol{m} \cdot \boldsymbol{e}_{x_i}$. Similarly, for $\mathcal{L}_{\mathbf{x}'}$, we have $\boldsymbol{m} \cdot \boldsymbol{e}_{x_i} \leq \boldsymbol{m} \cdot \boldsymbol{e}_{y_i}$. So when training model with loss function $\mathcal{L}_{\mathrm{MLM}} = \mathcal{L}_{\mathbf{x}} + \mathcal{L}_{\mathbf{x}'}$, model will be optimized to learn $\boldsymbol{m} \cdot \boldsymbol{e}_{x_i} = \boldsymbol{m} \cdot \boldsymbol{e}_{y_i}$. This equation can't ensure $\boldsymbol{e}_{x_i} = \boldsymbol{e}_{y_i}$ but $\boldsymbol{e}_{x_i} \sim \boldsymbol{e}_{y_i}$ to some extent. For standard masked language modeling, there is a probability that the original token will not be masked and we use $\boldsymbol{c}_{x_i}^{src}$ to represent the hidden state of the last layer, which is the contextualized embedding of token x_i. So we have $\boldsymbol{c}_{x_i}^{src} \cdot \boldsymbol{e}_k \leq \boldsymbol{c}_{x_i}^{src} \cdot \boldsymbol{e}_{x_i}$, $\forall k \in |\mathcal{V}|$. Obviously, $\boldsymbol{e}_{x_i} \sim \boldsymbol{c}_{x_i}^{src}$. Similarly, if we consider target language sentence $\mathbf{y} = \langle y_1, \cdots, y_{i-1}, y_i, y_{i+1}, \cdots, y_n \rangle$, we have $\boldsymbol{e}_{y_i} \sim \boldsymbol{c}_{y_i}^{tgt}$. So training model with masked language modeling on source-language,target-language and source-target code-switched sentences is approximately optimizing:

$$\boldsymbol{c}_{x_i}^{src} \sim \boldsymbol{e}_{x_i} \sim \boldsymbol{e}_{y_i} \sim \boldsymbol{c}_{y_i}^{tgt}$$