A Data and Training details

The training data consists of 1.8M sentence pairs selected from ZH⇒EN task of NIST08 Open Machine Translation Campaign [6] with 40.1M Chinese words and 48.3M English words respectively. The development set is chosen as NIST02, and the test set is NIST05. For DE⇒EN task, we adopt the WMT2016 news translation task. [7] Training data consists of 4.5M sentence pairs with 108M German words and 115M English words. The development set is chosen as newstest2009. To make NMT models capable of open-vocabulary translation, all the datasets are pre-processed by Byte Pair Encoding (BPE) (Sennrich et al., 2015) with joint 32K merging operations. [8]

We implemented the proposed alignment induction methods on Transformer (Vaswani et al., 2017), since it is the most popular NMT model nowadays and has multiple layer architecture for investigating how attention based method performs. For training Moses, we use all 1.8M sentences from the corpus, and we train a 4-gram language model based on the target side of its training data. For training both NMT models, only the sentences of length up to 256 tokens are used, with no more than 2^{15} tokens in a batch. The dimension of both word embeddings and hidden states are 512. Both encoder and decoder have 6 layers by default, and adopt multi-head attention with 8 heads. The beam size for decoding is 4, and the loss function is optimized by Adam (Kingma and Ba, 2014), where $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. Particularly, for the explicit alignment model, the alignment reference is produced by Fast Align.

B Pointwise Mutual Information among Words

Pointwise mutual information (PMI) measures the relevance of two discrete random variables, which is defined as

$$\begin{aligned}\mathrm{PMI}(\mu,\nu) &= \log \frac{P(\mu,\nu)}{P(\mu)\cdot P(\nu)} \\ &= \log Z + \log \frac{C(\mu,\nu)}{C(\mu)\cdot C(\nu)},\end{aligned} \tag{12}$$

where $C(\mu,\nu)$ is a function for counting occurrence of the pair (μ,ν) according to different scenarios, and Z is the normalizer, i.e. the total number of all possible (μ,ν) pairs. In this paper, we define two types of PMI according to different definitions of $C(\mu,\nu)$ in the two scenarios as follows.

PMI on Bilingual Data In this scenario, a set of bilingual sentences are given. For a given bilingual sentence $\langle \mathbf{x}, \mathbf{y}\rangle$, $C(\mathrm{y}_i, \mathrm{x}_j)$ is added by one if both $\mathrm{y}_i \in \mathbf{y}$ and $\mathrm{x}_j \in \mathbf{x}$.

PMI between a Word and Its History Word on Monolingual Data In this scenario, a set of monolingual sentences are given. For a given monolingual sentence $\mathbf{y}$, $C(\mathrm{y}_k, \mathrm{y}_i)$ is added by one if $\mathrm{y}_k \in \mathbf{y}$ and $\mathrm{y}_i \in \mathbf{y}$ with $k < i$.

C Different Margins for Dividing CFS and CFT

ϵ	Target Words	AER	RER	%
0	CFS	32.97	34.12	72.20
	CFT	63.28	37.90	27.80
10^{-4}	CFS	30.50	29.68	65.61
	CFT	62.91	33.35	24.66
10^{-3}	CFS	29.21	26.97	60.40
	CFT	63.29	30.47	22.04
10^{-2}	CFS	27.00	22.32	51.53
	CFT	64.22	26.17	17.56
10^{-1}	CFS	21.38	13.97	34.85
	CFT	64.13	21.58	10.39

* Results are measured on Transformer-L6 in ZH⇒EN task.

Table 8: AER and real decoding translation error rate (RER) under different partitions of CFS and CFT.

In equation 11, different margins will partition CFS and CFT differently. As growing of the margin ϵ, the partition of CFS and CFT becomes more confident. As shown in Table 8, both more confident CFS and CFT words can achieve lower real decoding translation error rate as the margin enlarging. But the alignment quality is only better for more confident CFS words instead of CFT words.

D Alignment Label Tool

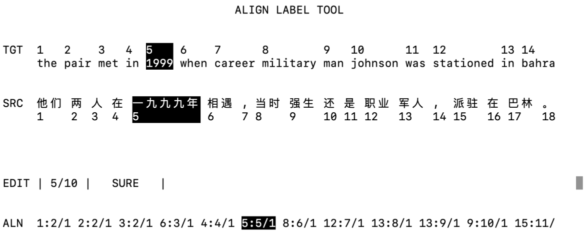
To measure the alignment performance in real translation, it is badly in need of an effective annotation tool for human to label ground-truth alignment between each translated target sentence and the source sentence. To this end, we

[6] LDC2000T50, LDC2002L27, LDC2002T01, LDC2002E18, LDC2003E07, LDC2003E14, LDC2003T17, LDC2004T07

[7] WMT14: http://www.statmt.org/wmt16/

[8] Throughout this paper, both BLEU and AER are evaluated after restoring BPE. Particularly for restoring BPE before acquiring alignment via MAP strategy in equation 3, $\boldsymbol{\alpha}$ is firstly merged along target tokens by averaging then merged along source tokens by summation.

develop an easy-to-use tool as shown in Figure 5 based on *curses* [9] and *python* to visualize and label the alignment between parallel sentences in command line interface. This software can be acquired from https://github.com/znculee/align-label-tool.



```
                                  ALIGN LABEL TOOL

TGT  1   2    3   4  5    6    7      8        9   10      11  12        13 14
     the pair met in 1999 when career military man johnson was stationed in bahra

SRC  他们 两 人 在 一九九九年 相遇 ，当时 强生 还 是 职业 军人 ， 派驻 在 巴林 。
     1    2  3  4  5          6    7 8    9    10 11 12   13   14 15   16 17   18


EDIT | 5/10 |   SURE   |

ALN  1:2/1 2:2/1 3:2/1 6:3/1 4:4/1 5:5/1 8:6/1 12:7/1 13:8/1 13:9/1 9:10/1 15:11/
```

Figure 5: Demo of word alignment labeling tool

[9] https://docs.python.org/3/howto/curses.html