Generating Diverse Translation by Manipulating Multi-Head Attention

Zewei Sun, Shujian Huang, Hao-Ran Wei, Xin-yu Dai, Jiajun Chen
State Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
sunzw@smail.nju.edu.cn, whr94621@foxmail.com
{huangsj, daixinyu, chenjj}@nju.edu.cn

Abstract
Transformer model (Vaswani et al. 2017) has been widely used in machine translation tasks and obtained state-of-the-art results. In this paper, we report an interesting phenomenon in its encoder-decoder multi-head attention: different attention heads of the final decoder layer align to different word translation candidates. We empirically verify this discovery and propose a method to generate diverse translations by manipulating heads. Furthermore, we make use of these diverse translations with the back-translation technique for better data augmentation. Experiment results show that our method generates diverse translations without a severe drop in translation quality. Experiments also show that back-translation with these diverse translations could bring a significant improvement in performance on translation tasks. An auxiliary experiment of conversation response generation task proves the effect of diversity as well.

Introduction
In recent years, neural machine translation (NMT) has shown its ability to produce precise and fluent translations (Sutskever, Vinyals, and Le 2014; Bahdanau, Cho, and Bengio 2015; Luong, Pham, and Manning 2015). More and more novel network structures has been proposed (Barone et al. 2017; Gehring et al. 2017; Vaswani et al. 2017), among which Transformer (Vaswani et al. 2017) achieves the best results. The main differences between Transformer and other translation models are: i) self-attention architecture, ii) multi-head attention mechanism. We focus on the second one in this paper.

Intuitively, the attention mechanism in traditional attention-based sequence-to-sequence models plays the role of choosing the next source word to be translated, which could be seen as an alignment between source and target words (Bahdanau, Cho, and Bengio 2015; Luong, Pham, and Manning 2015). However, how multi-head attention works seems unclear.

In this paper, we report an interesting phenomenon in Transformer: in the final layer of its decoder, each individual encoder-decoder attention head dispersedly aligns to a specific source word which is highly likely to be translated next. In other words, multi-head attention actually learns multiple alignment choices. Further, by means of picking different attention heads, we can precisely control the following word generation. We verify this characteristic by a series of statistical study afterwards.

Straightway, we consider taking advantage of this intrinsic characteristic to generate diverse translations due to the multiple generation candidates. Natural language can be diversely translated through different syntax structures or word orders. However, it has been well recognized that NMT system severely lacks translation diversity, distinguished from human beings (He, Haffari, and Norouzi 2018; Ott et al. 2018; Edunov et al. 2018). We try to tackle this issue with a new method based on our observation.

There have been a few works attempting to generate more diverse translation, which can be roughly divided into two categories. The first category tries to encourage diversity during beam search by adding some regularization items (Li, Monroe, and Jurafsky 2016; Vijayakumar et al. 2018). However, these methods actually fail to produce satisfied diversity in our re-implementation experiments. The other category tends to augment diversity by introducing latent variables (He, Haffari, and Norouzi 2018; Shen et al. 2019). However, they heavily increase training barrier and lack interpretability.

Different from them, we make use of the diverse factors we observe inside the model structure, which is more lightweight and interpretable. Since multi-head attention has the potential to identify different translation candidates, we propose a method to manipulate it to generate diverse translations. Our method works simply but more effectively than previous works, bringing in no extra parameter or regularization item. Furthermore, we propose to combine diverse translations with the back-translation technique for better data augmentation.

Experiment results show that the proposed method could generate diverse translations without a severe drop in translation performance. Besides, improvements could be achieved by employing our more diverse back-translation results for machine translation. An auxiliary experiment of conversation response generation proves the effect of diversity as well.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
Figure 1: Multi-Head Attention consists of several attention heads running in parallel.

Background
Transformer (Vaswani et al. 2017) architecture adopts the encoder-decoder structure, utilizing self-attention instead of recurrent or convolutional networks. The encoder iteratively processes its hidden representation through 6 layers of self-attention and feed-forward network, coupled with layer normalization and residual connection. Afterwards, the decoder takes a similar circuit with injected by an encoder-decoder attention layer between self-attention and feed-forward components. An important difference from previous models is that Transformer turns all the attention mechanism into a multi-head version.

Instead of performing a single attention function with $d$-dimensional keys, values and queries, multi-head attention projects them into $H$ different sub-components. After calculating attention for every sub-component, each yielding a $d/H$-dimensional output context, these context vectors are concatenated and projected, resulting in the final context, as depicted in Figure 1. Specifically:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)W^o \quad (1)$$

$$\text{head}_i = \text{Attention}(QW^Q_i, KW^K_i, VW^V_i) \quad (2)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

To complete the decoding part, the model uses learned linear transformation and softmax function to convert the decoder output to next-token probabilities. The embedding and final transformation parameters are shared mutually.

Analysis of Multi-Head Attention
Each Head Indicates an Alignment
Previous works show that multi-head attention plays a key role in the significant improvement of translation performance (Vaswani et al. 2017; Chen et al. 2018). However, not much observation was made on its inside pattern. We visualize the multi-head attention to see whether different heads play different roles. After observing plenty of examples, we find that at every decoding step, all the source words that are identified by heads are highly likely to be translated next. In other words, each head aligns to a source word candidate.

| Rank | NLL | Head | NLL |
|------|-----|------|-----|
| 1    | 0.59| 0    | 3.05|
| 3    | 4.12| 1    | 5.40|
| 5    | 5.19| 2    | 4.69|
| **Head-Average** | **5.27** | **3** | **5.04** |
| 6    | 5.53| 4    | 3.33|
| 10   | 6.40| 5    | 5.72|
| 100  | 9.47| 6    | 5.68|
| 1000 | 11.54| 7    | 5.32|
| Random | 13.57 | Random | 13.57 |

Table 1: Negative log-likelihood of attention heads and words ranked $R^{th}$ on average.

We do the statistic study to verify this observation on NIST MT03 dataset (see Datasets in Experiment). At each timestep, we pick $H$ referred source words (may overlap) that $H$ heads correspond to. “Referred source word” means the source word with max attention for each head. Then we translate these referred source words into target language with baseline model and name these translations “referred target words”. We count the number of times these “referred target words” appear at different rankings of the softmax probability and plot them in Figure 2. We can see that the vast majority of heads align to the most possible words.

Also, we collect the negative log-likelihood (NLL) of these referred target words to see whether they really have high generation probability. To make a comparison, we list the average NLL of words ranked $R^{th}$ as well. The results in Table 1 verify our assumption. The chosen words are ranked around $5^{th}$ on average, which implies they are indeed quite possible to be selected at each decoding step.

Each Head Determines a Generation Word
Furthermore, we can control the next word generation by choosing the corresponding source word by choosing different heads. As presented in Table 2, the model has translated “he said :” and waits for the following context. At this step, different heads refer to several source words. From Figure 3, we can see that head 4,5,6 refer to 以来 (since), 下降 (decline), 出口 (exports), respectively. We control the model to generate the specific word by selecting the corresponding

![Figure 2](image)

Figure 2: “Referred target words” ranking counts of top 100. The vast majority of heads align to the most possible words.
Source | 他说，去年九月以来出口下降导致印度经济恶化。
Reference | he said: the drop in exports has caused India’s economy deterioration since September last year.
Translated (since) | he said: since September last year, the decline in exports has led to a deterioration in India’s economy.
Translated (decline) | he said: the decline in exports has caused India’s economy to deteriorate since September last year.
Translated (exports) | he said: exports have declined since September last year, causing India’s economy to deteriorate.

Table 2: The model has translated “he said:” and waits for the following context. Different heads (refered to different candidates) determine different following generation.

Diversity-Encouraged Generation

Since we have confirmed multiple head alignments can be utilized, it is natural for us to sample different heads at every timestep, so that diverse word candidates can be generated. However, we found that it will badly harm the translation quality if sampling everywhere. So we propose a sample policy to balance quality and diversity.

As stated in Algorithm 1 at every decoding step $t$, we denote $att_h^t$ as the attention of $head_h$ from target side hidden state $s_t$ to source side word $src_i$. The most possible candidate and copying its attention weights to the other $H-1$ heads. In this way, we indeed obtain different translation results with expected translation candidates (see three translation outputs in Table 2).

Intuitively, we can utilize these characteristics to generate diverse translations by picking different candidates to change the word choices or the sentence structure. More importantly, the diversity is from an interpretable mechanism rather than an abstract latent variable like previous works.

Algorithm 1 Sample Policy

Input: The source sentence length $T$, a hyper parameters $K$, the head number $H$, a counting array $[n_0, ..., n_i, ..., n_{T-1}]$

Output: Adjusted attention

1: for $t$ in decoding timesteps do
2: for $i$ in range($T$) do
3: $n_i = 0$
4: end for
5: calculate $att_h^t$, $i \in [0, T), h \in [0, H)$
6: for $h$ in range($H$) do
7: $candidate_h^t = \arg \max_i att_h^t$
8: $n_{candidate_h^t} + = 1$
9: end for
10: if $\max(n) \leq K$ then
11: head = sample$[0, H)$
12: for all $h$ do
13: $att_h^t = att_{head}^t$
14: end for
15: end if
16: end for

We denote an array of $[n_0, ..., n_i, ..., n_{T-1}]$ as the number of times that word $i$ is chosen from equation 4, where $T$ is the length of source sentence. Obviously:

$$\sum_{i=0}^{T-1} n_i = H$$

(5)

where $H$ is the number of heads. Diverse translations are generated when multiple candidates are offered. In other words, not all heads focus on the same source word. Therefore, we define a confusing condition when:

$$\max(n_i) \leq K$$

(6)

where $K$ is a hyper-parameter. Confusing condition means “referred words” are disperse and multiple candidates can be accepted. Under confusing condition, we sample one of the heads as attention and force other heads to be the same.
Otherwise, the decoding step remains unchanged. If \( K = 0 \), the model is the same as the original version. If \( K = H \), the model samples at every step. For the sake of balance of quality and diversity, we may choose different \( K \) in different conditions. We do the whole decoding for \( M \) times and pick the most possible output in the beam every time. In this paper, we let \( M = 5 \).

Our another contribution is adopting the method with back-translation technique as data augmentation. Back-translation has been proved helpful for neural machine translation (Sennrich, Haddow, and Birch 2016a, Poncelas et al. 2018). However, the lack of diversity restricts its effect (Edunov et al. 2018). We provide a new scheme for back-translation with diverse corpus generated by our method and gain improvement.

**Experiment**

Our translation experiments include two parts: diverse translation and diverse back-translation. In addition, a conversation response generation experiment is also performed as auxiliary evidence.

**Setup**

**Datasets** We choose five datasets as our experiment corpus.

- NIST Chinese-to-English (NIST Zh-En). The training data consists of 1.34 million sentence pairs extracted from LDC corpus. We use MT03 as the development set, MT04, MT05, MT06 as the test sets.
- WMT14 English-to-German (WMT En-De). The training data consists of 4.5 million sentence pairs from WMT14 news translation task. We use newstest2013 as the development set and newstest2014 as the test set.
- WMT16 English-to-Romanian (WMT En-Ro). The training data consists of 0.6 million sentence pairs from WMT16 news translation task. We use newstest2015 as the development set and newstest2016 as the test set.
- Monolingual English corpus (for back-translation) from IWSLT17 Chinese-to-English (IWSLT Zh-En). The training data consists of 0.2 million sentence pairs from IWSLT17 spoken language translation task. We used dev2010 and tst2010 as the development set and tst2011 as the test set.
- Short Text Conversation (STC) (Shang, Lu, and Li 2015). The corpus contains about 4.4 million Chinese post-response sentence pairs crawled from Weibo, built for single turn conversation tasks. We remove sentence pairs that are exactly the same between two sides. For the test set, we extract 3000 post sentences that have 10 responses in the corpus, forming 10 references. The develop set is made up similarly.

For NIST Zh-En, we use BPE (Sennrich, Haddow, and Birch 2016b) with 30K merge operations on both sides. For En-De and En-Ro, we also apply BPE to segment sentences and limit the vocabulary size to 32K. We filter out sentence pairs whose source or target side contains more than 100 words for Zh-En and En-Ro sets. For STC corpus, we also apply BPE and keep a vocabulary size to 36K. All the out-of-vocabulary words are mapped to a distinct token \(<\text{UNK}>\).

**Experiment Settings** Without extra statement, we follow the Transformer base v1 setting\(^1\) with 6 layers in encoder and 2 layers in decoder\(^2\) 512 hidden units, 8 heads in multi-head attention and 2048 hidden units in feed-forward layers. Parameters are optimized using Adam optimizer (Kingma and Ba 2015), with \( \beta_1 = 0.9, \beta_2 = 0.98, \) and \( \epsilon = 10^{-9} \). The learning rate is scheduled according to the method proposed in Vaswani et al. (2017), with \textit{warmup_steps} = 8000. Label smoothing (Szegedy et al. 2016) of value \( 0.1 \) is also adopted. For \( K \), we do not observe diversity enhancement when \( K \) is too small like \( K = 1, 2 \). And conditions of \( K = 6, 7 \) are very similar to \( K = 8 \). Hence we use \( K = 3, 4, 5, 6 \) as comparisons.

**Diverse Translation**

**Comparing Objects** We compare our models with original beam search (as Baseline) and sampling from the probability distribution (as Multinomial Sampling). Besides, we compare our methods with a few previous works:

- Li, Monroe, and Jurafsky (2016) propose a decoding trick to penalize hypotheses that are siblings (expansions of the same parent node) in the beam search to increase the translation diversity.
- Vijayakumar et al. (2018) add a regularization item in beam search to penalize the same word generation.
- Shen et al. (2019) and He, Haffari, and Norouzi (2018) use multiple decoders as mixture of experts to increase diversity by manipulating latent variables. Considering their similarity, we choose Shen et al. (2019) since they report better results. We re-implement the model with Transformer architecture and choose the hMup (online-shared) version since the authors recommend it.

**Metrics** We evaluate our method from both diversity and quality. For diversity, we adopt average pair-wise BLEU of outputs (denoted as \( \text{pwb} \)) to measure the difference among translations like previous work. For quality, we use BLEU with the references (denoted as \( \text{rfb} \)). Lower \( \text{pwb} \) and higher \( \text{rfb} \) mean better results. In this paper, the reference BLEU of Baseline is the highest score in the beam while the other groups take the average reference BLEU of \( M \) outputs. And to synthetically evaluate the performance, we propose an overall index: Diversity Enhancement per Quality (denoted as DEQ). Specifically:

\[
\text{DEQ} = \frac{\text{pwb}^* - \text{pwb}}{\text{rfb}^* - \text{rfb}}
\]

where \( \text{pwb} \) and \( \text{pwb}^* \) refer to pair-wise BLEU of the evaluated system and baseline respectively, \( \text{rfb} \) and \( \text{rfb}^* \) refer to

\(^1\)https://github.com/tensorflow/tensorflow/blob/v1.3.0/tensor2tensor/models/transformer.py

\(^2\)We check different decoder layer numbers settings and find less-decoder-layers Transformer shows comparable performance with original six-layers-decoder Transformer while it is much easier to manipulate and faster to decode. The diversity enhancement is also more significant. We present detailed research in Appendix, which is highly recommended to refer to.
| Model                      | MT03 (dev) rfb↑  | MT04 pwb↓  | MT05 rfb↑  | MT06 pwb↓  | MT07 rfb↑  | MT08 pwb↓  | Average DEQ↑ |
|---------------------------|------------------|------------|------------|------------|------------|------------|-------------|
| Baseline                  | 45.64 84.63      | 47.25 84.62 | 43.45 84.78 | 42.26 82.46 | 44.32 83.95 | 84.78 42.26 |            |
| Multinomial Sampling       | 21.75 11.29      | 22.19 11.42 | 20.54 11.08 | 19.12 9.67  | 20.62 10.72 | 11.08 20.54 | 3.09        |
| (Li, Monroe, and Jurafsky 2016) | 44.63 80.92      | 45.81 81.33 | 42.86 81.28 | 40.87 78.11 | 43.18 80.24 | 81.33 42.86 |            |
| [Vijayakumar et al. 2018]  | 40.38 39.55      | 41.99 60.11 | 39.46 59.56 | 37.28 54.54 | 39.58 58.07 | 60.11 39.46 |            |
| (Shen et al. 2019)        | 40.59 62.24      | 41.55 62.68 | 38.51 61.37 | 35.57 58.04 | 38.54 60.70 | 62.68 38.51 |            |
| Sample K = 3              | 43.73 66.48      | 45.38 67.82 | 42.43 65.80 | 40.18 64.93 | 42.66 66.18 | 66.48 43.73 |            |
| Sample K = 4              | 40.88 51.26      | 42.50 53.63 | 39.18 51.07 | 37.73 50.28 | 39.80 51.66 | 51.26 40.88 | 7.14        |
| Sample K = 5              | 38.60 43.64      | 40.21 45.69 | 37.05 43.14 | 35.45 42.38 | 37.57 43.74 | 43.64 38.60 | 9.67        |
| Sample K = 8              | 36.68 38.29      | 38.03 40.02 | 34.65 37.30 | 32.93 36.15 | 35.20 37.82 | 38.29 36.68 | 5.06        |

Table 3: Pair-wise BLEU and Reference BLEU in Zh2En experiments.

| Model                      | MT03 (dev) rfb↑  | MT04 pwb↓  | MT05 rfb↑  | MT06 pwb↓  | MT07 rfb↑  | MT08 pwb↓  | Average DEQ↑ |
|---------------------------|------------------|------------|------------|------------|------------|------------|-------------|
| Baseline                  | 26.31 80.41      | -          |            |            |            |            |             |
| Multinomial Sampling       | 11.99 12.84      | 4.72       |            |            |            |            |             |
| (Li, Monroe, and Jurafsky 2016) | 25.27 78.57      | 1.77       |            |            |            |            |             |
| [Vijayakumar et al. 2018]  | 23.27 66.13      | 4.70       |            |            |            |            |             |
| (Shen et al. 2019)        | 23.22 68.03      | 4.01       |            |            |            |            |             |
| Sample K = 3              | 25.62 78.96      | 2.10       |            |            |            |            |             |
| Sample K = 4              | 24.26 62.04      | 8.96       |            |            |            |            |             |
| Sample K = 5              | 22.62 50.14      | 8.20       |            |            |            |            |             |
| Sample K = 8              | 19.76 38.36      | 6.42       |            |            |            |            |             |

Table 4: Pair-wise BLEU and Reference BLEU in En2De experiments.

| Model                      | MT03 (dev) rfb↑  | MT04 pwb↓  | MT05 rfb↑  | MT06 pwb↓  | MT07 rfb↑  | MT08 pwb↓  | Average DEQ↑ |
|---------------------------|------------------|------------|------------|------------|------------|------------|-------------|
| Baseline                  | 31.76 81.29      | -          |            |            |            |            |             |
| Multinomial Sampling       | 18.85 20.82      | 4.68       |            |            |            |            |             |
| (Li, Monroe, and Jurafsky 2016) | 31.02 78.42      | 3.88       |            |            |            |            |             |
| [Vijayakumar et al. 2018]  | 28.91 69.67      | 4.08       |            |            |            |            |             |
| (Shen et al. 2019)        | 31.07 85.71      | -6.04      |            |            |            |            |             |
| Sample K = 3              | 31.33 82.41      | 2.60       |            |            |            |            |             |
| Sample K = 4              | 30.06 71.12      | 5.98       |            |            |            |            |             |
| Sample K = 5              | 27.89 59.42      | 5.65       |            |            |            |            |             |
| Sample K = 8              | 26.43 50.56      | 5.77       |            |            |            |            |             |

Table 5: Pair-wise BLEU and Reference BLEU in En2Ro experiments.

Result From Table 3 in Zh2En experiment, we can see that traditional beam search translations severely lack diversity while multinomial sampling extremely harms translation quality. Li, Monroe, and Jurafsky (2016) bring very limited enhancement in diversity, failing to achieve the goal. Vijayakumar et al. (2018) and Shen et al. (2019) show the ability to produce diversity, but our method (K = 4) attains more significant diversity as well as better quality comparing with them. And K = 3 achieve the highest DEQ, gaining the most satisfactory result. Also, unlike Shen et al. (2019), our work needs no extra training or extra parameters. Furthermore, the diversity can be well interpreted and does not rely on an abstract latent variable. See Table 6 for a case (more cases are in Appendix). We can reach the similar conclusion in En2De and En2Ro experiments from Table 4, 5.

Besides, to exclude the possibility that randomly interfering causes the effect, we compare with the sets with noise. We add noise to the translations of baseline model to generate different outputs. Specifically, for each sentence, we replace one of its words with ‘<UNK>’ with probability p and randomly swap two words with probability p as well. And we make multiple experiment sets by controlling p. See Figure 4 at the same level of pair-wise BLEU, our method maintains much higher reference BLEU, which means our method improves diversity through seeking diverse translations rather than just generating randomly.

For K (see Figure 4 again), as expected, as K grows (sample more), the diversity increases (pair-wise BLEU decreases) while the quality decreases (reference BLEU decreases). And previous works all lie on the top left of the curve of K. What’s more, we can choose different K to diversely balance the diversity and the quality depending on our needs. We make the trade-off more continuous. Some may not be satisfied with the sacrifice of the reference BLEU. But considering the calculation of BLEU is based on n-gram rather than semantic similarity, we regard it as a normal phenomenon. After all, if we want to obtain sentences with different grammar structure or word order, the overlap of n-gram will inevitably decrease to some extent.
as back-translation, denoted as Beam-5. According to Edunov et al. (2018), unrestricted sampling from the model distribution yields the best performance as our work adds generation diversity and maintains the quality simultaneously.

We do experiments under conditions with and without additional monolingual data.

**Self Back-Translation** Firstly, we focus on the condition where original training data is repeatedly used by back-translation. When translating language pair $f$ to $e$, for each target sentence $e$, we get $M$ translations with a reverse translation model. We combine those translations with $e$ as synthetic sentence pairs and add them to the training data. As previously stated, we let $M = 5$. Experiments are conducted on Zh-En NIST dataset.

See Table 7 and 8, all of our experiment sets report better results, among which, the best set of $K = 3$ in Zh2En experiments yields 1.82 improvement and the best set of $K = 4$ in En2Zh experiments yields 0.82 improvement.

**Utilizing Additional Monolingual Data** Secondly, we evaluate our method with additional monolingual data. We select one side of parallel data from IWSLT17 as monolingual data. We use the same method to generate synthetic sentence pairs. Then we train our model on the mixture of original NIST dataset and the synthetic dataset.

See Table 9, experiment results show that our algorithm brings the most significant improvement for translation performance as our work adds generation diversity and maintains the quality simultaneously.

**Conversation Response Generation** Responses generated by neural conversational models tend to lack informativeness and diversity (Li et al. 2016; Shao et al. 2017; Baheti et al. 2018; Zhang et al. 2018). Therefore, we try to ease this issue by utilizing our method on Conversation Response Generation tasks. Still, we perform decoding for $M$ times and pick the $N^{th}$ output in the beam for the $N^{th}$ group ($N \in [1, M]$).
mar et al. (2018) adds a regularization item in beam search to increase the translation diversity. Vijayakumar siblings (expansions of the same parent node) in the forward some related methods. Li, Monroe, and Jurafsky machine translation. In recent years, a few works put Lack of diversity has been a disturbing problem for neu-source side. In contrast, our method brings in randomness, safe outputs like “I don’t know” or simply copying from the original post. We classify responses into two groups, informative and un-informative. Uninformative means the safe answer like “I don’t know” or simply copying from the original post. We then calculate the proportion of the informative groups. For diversity, pair-wise BLEU maintains used.

Metrics Since the responses of human conversation can be pretty subjective, which is hard to evaluate automatically. Hence, except for reference BLEU, we also measure the response quality by human evaluation through three indexes: relevance, fluency and informativeness. Relevance reveals how much the responses match the expectation of the question. Fluency means to what extent the translation is well-formed grammatically. Both of them are scored from 1 to 5. Informativeness measures the degree of meaningfulness. We classify responses into two groups, informative and un-informative. Uninformative means the safe answer like “I don’t know” or simply copying from the original post. We then calculate the proportion of the informative groups. For diversity, pair-wise BLEU maintains used.

Results In Table 10, we compare our method with basic Seq2Seq model and Li et al. (2016), which use Maximum Mutual Information (MMI) as the objective function (MMI-antiLM version). On one hand, our method achieves significant improvement in generation diversity. On the other hand, the quality including relevance, fluency and informativeness all rise to some degree. After looking into cases, we suppose it is because original Seq2Seq model tends to generate safe outputs like “I don’t know” or simply copying from the source side. In contrast, our method brings in randomness, reaching broader generation space.

Related Work

Lack of diversity has been a disturbing problem for neural machine translation. In recent years, a few works put forward some related methods. Li, Monroe, and Jurafsky (2016) proposes a decoding trick to penalize hypotheses that are siblings (expansions of the same parent node) in the beam search to increase the translation diversity. Vijayakumar et al. (2018) adds a regularization item in beam search to penalize the same word generation. He, Haffari, and Norouzi (2018) and Shen et al. (2019) use multiple decoders as different components, trying to control the generation by different latent variables. Basically, there are two categories: either to add diverse regularization in beam search or to utilize latent variables. Our method achieves better results than both two categories. And specifically, compared with the latter class, our work needs no extra training or extra parameters. Besides, it is hard to tell what the latent variables exactly represent and why they differ while our method shows a clear explanation that heads align to word candidates.

Apart from machine translation, there are also other works concerning generation diversity, including Visual Question Generation (Jain, Zhang, and Schwingle 2017), Conversational Response Generation (Li et al. 2016; Shao et al. 2017; Babeti et al. 2018; Zhang et al. 2018), Paraphrase (Gupta et al. 2018; Xu et al. 2018b), Summarization (Nema et al. 2017) and Text Generation (Guan et al. 2018; Xu et al. 2018a).

As for multi-head attention, Strubell et al. (2018) employ different heads to capture different linguistic features. Tu et al. (2018) introduce disagreement regularization to encourage diversity among attention heads. Li et al. (2019) propose to aggregate information captured by different heads. Yang et al. (2019) model the interactions among attention heads. Raganato and Tiedemann (2018) do an analysis of encoder representation and find there exists dependency relations, syntactic and semantic connections across layers.

**Conclusion**

In this paper, we discover an internal characteristic of Transformer encoder-decoder multi-head attention that each head aligns to a source word which is a possible candidate to be translated. We take advantage of this phenomenon to generate diverse translations by manipulating heads in particular conditions. Experiments show that our algorithm outperforms previous work and obtain the most satisfactory result of quality and diversity. Besides, the multiple trade-off setting can be adopted diversely depending on different needs. Finally, applications on back-translation as data augmentation and conversation response significantly improve the performance, proving our method effective.

**Acknowledgement**

Shujian Huang is the corresponding author. This work is supported by the National Key R&D Program of China (No. U1836221, 61772261), the Jiangsu Provincial Research Foundation for Basic Research (No. BK20170074).

---

Table 8: En2Zh translation experiments with back-translation of original training data.

| Model  | MT03  | MT04  | MT05  | MT06  | Average |
|-------|-------|-------|-------|-------|---------|
| Baseline | 22.75 | 22.33 | 20.35 | 21.35 | 21.34   |
| Beam-5  | 23.73 | 21.69 | 20.61 | 22.33 | 21.54   |
| Sampling | 23.69 | 22.78 | 20.85 | 22.34 | 21.99   |
| K = 3   | **24.21** | 22.23 | 20.65 | **22.52** | 21.80   |
| K = 4   | **24.01** | **23.15** | **21.04** | 22.30 | **22.16** |
| K = 5   | 23.76 | 21.93 | 20.57 | 22.50 | 21.67   |
| K = 8   | 23.93 | 21.66 | 20.72 | 22.23 | 21.54   |

Table 9: Zh2En translation experiments with back-translation of additional monolingual data.

| Model  | Zh2En |
|-------|-------|
| Baseline | 9.18 |
| Beam-5  | 13.06 |
| Sampling | 13.38 |
| K = 3   | **14.03** |
| K = 4   | 13.76 |
| K = 5   | 13.66 |
| K = 8   | 13.76 |

Table 10: Conversation Response Generation experiment results on STC dataset.

| Model  | BLEU ↑ | Rel ↑ | Flu ↑ | Inf ↑ | Div ↓ |
|-------|--------|-------|-------|-------|-------|
| Baseline | 13.06 | 2.45 | 4.72 | 0.604 | 52.94 |
| MMI     | 13.39 | 2.58 | 4.45 | 0.639 | 45.42 |
| K = 3   | **13.48** | **2.63** | **4.76** | **0.678** | **39.82** |
| K = 8   | 12.73 | 2.53 | 4.67 | 0.652 | 27.73 |
References

[2015] Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In ICLR.

[2018] Baheti, A.; Ritter, A.; Li, J.; and Dolan, W. B. 2018. Generating more interesting responses in neural conversation models with distributional constraints. In EMNLP.

[2017] Barone, A. V. M.; Helcl, J.; Sennrich, R.; Haddow, B.; and Birch, A. 2017. Deep architectures for neural machine translation. In WMT.

[2018] Chen, M. X.; Firat, O.; Bapna, A.; Johnson, M.; Macherey, W.; Foster, G.; Jones, L.; Parmar, N.; Schuster, M.; Chen, Z.; Wu, Y.; and Hughes, M. 2018. The best of both worlds: Combining recent advances in neural machine translation. In ACL.

[2018] Edunov, S.; Ott, M.; Auli, M.; and Grangier, D. 2018. Understanding back-translation at scale. In EMNLP.

[2017] Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; and Ranzato, M. 2017. A simple, effective approach to attention-based neural machine translation. In EMNLP.

[2018] Gupta, A.; Agarwal, A.; Singh, P.; and Rai, P. 2018. A deep generative framework for paraphrase generation. In AAAI.

[2018] Guu, K.; Hashimoto, T. B.; Oren, Y.; and Liang, P. S. 2018. Generating sentences by editing prototypes. Transactions of the Association for Computational Linguistics 6:437–450.

[2018] He, X.; Haffari, G.; and Norouzi, M. 2018. Sequence to sequence mixture model for diverse machine translation. In CoNLL.

[2017] Jain, U.; Zhang, Z.; and Schwing, A. G. 2017. Creativity: Generating diverse questions using variational autoencoders. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 5415–5424.

[2015] Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. In ICLR.

[2016] Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, W. B. 2016. A diversity-promoting objective function for neural conversation models. In HLT-NAACL.

[2019] Li, J.; Yang, B.; Dou, Z.-Y.; Wang, X.; Lyu, M. R.; and Tu, Z. 2019. Information aggregation for multi-head attention with routing-by-agreement. In NAACL.

[2016] Li, J.; Monroe, W.; and Jurafsky, D. 2016. A simple, fast diverse decoding algorithm for neural generation. CoRR abs/1611.08562.

[2015] Luong, T.; Pham, H. Q.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. In EMNLP.

[2017] Nema, P.; Khatra, M. M.; Laha, A.; and Ravindran, B. 2017. Diversity driven attention model for query-based abstractive summarization. In ACL.

[2018] Ott, M.; Auli, M.; Grangier, D.; and Ranzato, M. 2018. Analyzing uncertainty in neural machine translation. In ICML.

[2018] Poncelas, A.; Shterionov, D.; Way, A.; de Buyl Wenginger, G. M.; and Passban, P. 2018. Investigating back-translation in neural machine translation. CoRR abs/1804.06189.

[2018] Raganato, A., and Tiedemann, J. 2018. An analysis of encoder representations in transformer-based machine translation. In BlackboxNLP@EMNLP.

[2016a] Sennrich, R.; Haddow, B.; and Birch, A. 2016a. Improving neural machine translation models with monolingual data. In ACL.

[2016b] Sennrich, R.; Haddow, B.; and Birch, A. 2016b. Neural machine translation of rare words with subword units. In ACL.

[2015] Shang, L.; Lu, Z.; and Li, H. 2015. Neural responding machine for short-text conversation. In ACL.

[2017] Shao, Y.; Gouws, S.; Britz, D.; Goldie, A.; Strope, B.; and Kurzweil, R. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In EMNLP.

[2019] Shen, T.; Ott, M.; Auli, M.; and Ranzato, M. 2019. Mixture models for diverse machine translation: Tricks of the trade. In ICML.

[2018] Strubell, E.; Verga, P.; Andor, D.; Weiss, D. I.; and McCallum, A. 2018. Linguistically-informed self-attention for semantic role labeling. In EMNLP.

[2014] Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In NIPS.

[2016] Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2818–2826.

[2018] Tu, Z.; Yang, B.; Lyu, M. R.; and Zhang, T. 2018. Multi-head attention with disagreement regularization. In EMNLP.

[2017] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In NIPS.

[2018] Vijayakumar, A. K.; Cogswell, M.; Selvaraju, R. R.; and Tu, Z. 2018. Understanding back-translations at scale. In BlackboxNLP@EMNLP.

[2018a] Xu, J.; Ren, X.; Lin, J.; and Sun, X. 2018a. Diversity-promoting gan: A cross-entropy based generative adversarial network for diversified text generation. In EMNLP.

[2018b] Xu, Q.; Zhang, J.; Qu, L.; Xie, L.; and Nock, R. 2018b. D-page: Diverse paraphrase generation. CoRR abs/1808.04364.

[2019] Yang, B.; Wang, L.; Wong, D.; Chao, L. S.; and Tu, Z. 2019. Convolutional self-attention networks. In NAACL.

[2018] Zhang, Y.; Galley, M.; Gao, J.; Gan, Z.; Li, X.; Brockett, C.; and Dolan, W. B. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In NeurIPS.
Less-decoder-layers Transformer works fine as well

We check the different experimental groups of different decoder layer numbers to see whether the alignments exist widely. It turns out that less-decoder-layers Transformer (six-layers encoder unchanged) has this property as well. And owing to its shallower structure, we find the phenomenon more common and obvious. Simultaneously, it is much easier to control the word candidates by picking different heads from the experiment. For diversity, see Table 13 and Figure 6.

As for the performance, to our surprise, less-decoder-layers Transformer performs comparably with six-decoder-layers Transformer (see Table 11, 12). What’s more, less decoder layers bring higher speed (see Table 11).

Thus, for the sake of diversity significance and decoding speed, we choose the two-layers-decoder Transformer structure as our default setting unless stated otherwise.

More cases

We select some typical cases of diverse translation from Zh2En experiments (see Table 14).

| n | MT03 | MT04 | MT05 | MT06 | Average | Speed |
|---|------|------|------|------|---------|-------|
| 6 | 45.83| 46.66| 43.36| 42.17| 44.06   | 1 X   |
| 1 | 44.55| 45.73| 41.86| 40.53| 42.71   | 2.79 X|
| 2 | 45.64| 47.25| 43.45| 42.26| 44.32   | 2.34 X|
| 3 | 45.85| 46.99| 43.74| 42.77| 44.50   | 1.77 X|
| 4 | 45.29| 46.77| 43.71| 42.09| 44.19   | 1.58 X|
| 5 | 45.27| 46.04| 43.33| 43.06| 44.14   | 1.26 X|
| 8 | 44.81| 46.11| 42.24| 42.15| 43.50   | 0.69 X|
| 12| 45.34| 45.49| 43.09| 42.41| 43.66   | 0.67 X|

Table 11: BLEU and speed comparisons between different decoder layer numbers in Zh2En Experiments

| n_decoders | En2De | En2Ro |
|------------|-------|-------|
| two        | 26.31 | 31.76 |
| six        | 26.70 | 31.86 |

Table 12: BLEU comparisons between different decoder layer numbers in En2De and En2Ro Experiments
Table 13: Pair-wise BLEU and Reference BLEU in Zh2En experiments of different decoder layer numbers and different $K$ sets.

| n_d | Sample K | MT03 (dev) | MT04 | MT05 | MT06 | Average |
|-----|----------|------------|------|------|------|---------|
|     |          | rtb↑ | pwb↓ | rtb↑ | pwb↓ | rtb↑ | pwb↓ | rtb↑ | pwb↓ |
| Baseline | 44.55 | 83.38 | 45.73 | 83.68 | 41.86 | 84.28 | 40.53 | 81.32 | 42.71 | 83.09 |
| K = 3 | 39.15 | 54.91 | 40.18 | 56.29 | 37.02 | 53.58 | 35.47 | 54.44 | 37.56 | 54.44 |
| 1 | 33.86 | 40.96 | 35.39 | 43.09 | 32.11 | 40.07 | 30.38 | 39.58 | 32.63 | 40.91 |
| K = 5 | 30.89 | 33.31 | 32.37 | 35.42 | 29.26 | 33.01 | 27.66 | 32.43 | 29.76 | 33.62 |
| K = 8 | 26.94 | 26.63 | 28.99 | 28.97 | 25.57 | 26.99 | 24.41 | 26.62 | 26.32 | 27.53 |
| Baseline | 45.64 | 84.63 | 47.25 | 84.62 | 43.45 | 84.78 | 42.26 | 82.46 | 44.32 | 83.95 |
| K = 3 | 43.73 | 66.48 | 45.38 | 67.82 | 42.43 | 65.80 | 40.18 | 64.93 | 42.66 | 66.18 |
| K = 4 | 40.88 | 51.26 | 42.50 | 53.63 | 39.18 | 51.07 | 37.73 | 50.28 | 41.02 | 51.37 |
| K = 5 | 38.60 | 43.64 | 40.21 | 45.69 | 36.06 | 43.14 | 35.48 | 42.96 | 37.56 | 43.74 |
| K = 8 | 36.68 | 38.29 | 38.03 | 40.02 | 34.65 | 37.30 | 32.93 | 36.15 | 35.20 | 37.82 |
| Baseline | 45.29 | 84.35 | 46.77 | 84.62 | 43.71 | 84.70 | 42.09 | 82.25 | 44.19 | 83.86 |
| K = 3 | 44.25 | 72.03 | 46.02 | 72.81 | 42.96 | 70.74 | 41.45 | 70.06 | 43.48 | 71.20 |
| 4 | 42.91 | 59.93 | 44.59 | 61.15 | 41.50 | 58.77 | 40.07 | 58.34 | 42.05 | 59.42 |
| K = 5 | 41.90 | 53.15 | 43.64 | 54.92 | 40.36 | 52.71 | 39.07 | 51.87 | 41.02 | 53.17 |
| K = 8 | 40.92 | 48.34 | 43.31 | 54.02 | 36.06 | 43.14 | 35.48 | 42.96 | 37.56 | 43.74 |
| Baseline | 45.27 | 84.39 | 46.04 | 84.74 | 43.33 | 84.92 | 43.06 | 82.46 | 44.14 | 84.04 |
| K = 3 | 44.41 | 74.45 | 45.58 | 75.84 | 42.66 | 74.57 | 42.53 | 72.42 | 41.75 | 71.64 |
| 5 | 43.61 | 65.59 | 44.77 | 67.01 | 42.04 | 64.83 | 41.73 | 64.03 | 42.85 | 65.29 |
| K = 5 | 42.44 | 57.72 | 44.33 | 59.36 | 40.93 | 57.40 | 40.24 | 55.91 | 41.83 | 57.56 |
| K = 8 | 41.43 | 52.84 | 43.31 | 55.15 | 39.92 | 53.29 | 39.32 | 52.05 | 40.85 | 53.50 |
| Baseline | 45.83 | 84.24 | 46.66 | 84.43 | 43.36 | 84.71 | 42.17 | 82.16 | 44.06 | 83.77 |
| K = 3 | 44.76 | 74.36 | 45.99 | 75.42 | 42.85 | 73.65 | 41.61 | 72.65 | 43.48 | 73.91 |
| 6 | 43.87 | 66.50 | 45.50 | 68.26 | 42.08 | 65.89 | 41.25 | 65.51 | 42.94 | 66.55 |
| K = 5 | 43.65 | 62.53 | 45.13 | 64.89 | 41.73 | 62.32 | 40.83 | 61.61 | 42.56 | 62.94 |
| K = 8 | 42.98 | 59.58 | 44.61 | 62.22 | 41.31 | 59.61 | 40.69 | 59.05 | 42.20 | 60.29 |
| Baseline | 44.81 | 84.55 | 46.11 | 84.77 | 42.24 | 85.11 | 42.15 | 82.70 | 43.5 | 84.19 |
| K = 3 | 43.80 | 73.45 | 45.61 | 75.91 | 41.81 | 73.38 | 41.58 | 73.10 | 43.02 | 74.13 |
| 8 | 43.53 | 67.72 | 45.43 | 70.30 | 41.35 | 67.54 | 41.39 | 67.31 | 42.72 | 68.38 |
| K = 5 | 43.59 | 65.29 | 45.09 | 68.05 | 41.25 | 64.65 | 41.10 | 64.12 | 42.48 | 65.61 |
| K = 8 | 42.84 | 63.57 | 45.18 | 66.64 | 41.21 | 63.21 | 40.85 | 62.92 | 42.41 | 64.26 |
| Baseline | 45.34 | 84.51 | 45.49 | 84.34 | 43.09 | 84.86 | 42.41 | 82.15 | 43.66 | 83.78 |
| K = 3 | 44.84 | 75.90 | 45.33 | 77.66 | 42.47 | 76.00 | 42.32 | 74.97 | 43.37 | 76.21 |
| 12 | 44.77 | 72.38 | 45.21 | 73.85 | 42.33 | 72.37 | 42.02 | 71.21 | 43.19 | 72.48 |
| K = 5 | 44.57 | 70.53 | 45.19 | 72.51 | 42.24 | 70.72 | 42.06 | 69.71 | 43.16 | 70.98 |
| K = 8 | 44.73 | 70.43 | 45.18 | 72.40 | 42.30 | 69.95 | 41.91 | 68.85 | 43.13 | 70.40 |
1. he said, however, that after a US military training, the training of Iraqi troops will be fruitful.

2. He, however, believes that after US training for the US military over a long period of time, Iraq's training will be successful.

3. However, he believed that the Iraqi troops' training will be successful after a long period of training US troops will be successful.

4. He said, however, that after a military training, the training of Iraqi troops will be fruitful.

5. He, however, believes that after being trained for a long time by the US army, Iraqi forces will achieve good training results.

---

**Table 14: More Cases**