

Towards User-Driven Neural Machine Translation

Huan Lin^{1,2} Liang Yao³ Baosong Yang³ Dayiheng Liu³ Haibo Zhang³
Weihua Luo³ Degen Huang⁴ Jinsong Su^{1,2,5*}

¹School of Informatics, Xiamen University

²Institute of Artificial Intelligence, Xiamen University

³Alibaba Group ⁴Dalian University of Technology ⁵Pengcheng Lab, Shenzhen

huanlin@stu.xmu.edu.cn

{yaoliang.yl, yangbaosong.ybs, liudayiheng.ldyh, zhanhui.zhb}@alibaba-inc.com

weihua.luowh@alibaba-inc.com huangdg@dlut.edu.cn jssu@xmu.edu.cn

Abstract

A good translation should not only translate the original content semantically, but also incarnate personal traits of the original text. For a real-world neural machine translation (NMT) system, these **user traits** (e.g., topic preference, stylistic characteristics and expression habits) can be preserved in user behavior (e.g., historical inputs). However, current NMT systems marginally consider the user behavior due to: 1) the difficulty of modeling user portraits in zero-shot scenarios, and 2) the lack of user-behavior annotated parallel dataset. To fill this gap, we introduce a novel framework called **user-driven NMT**. Specifically, a **cache-based module** and a **user-driven contrastive learning method** are proposed to offer NMT the ability to capture potential user traits from their historical inputs under a zero-shot learning fashion. Furthermore, we contribute the first Chinese-English parallel corpus annotated with user behavior called **UDT-Corpus**. Experimental results confirm that the proposed user-driven NMT can generate user-specific translations.¹

1 Introduction

In recent years, neural machine translation (NMT) models (Sutskever et al., 2014; Luong et al., 2015; Vaswani et al., 2017) have shown promising quality and thus increasingly attracted users. When drawing on a translation system, every user has his own traits, including topic preference, stylistic characteristics, and expression habits, which can be implicitly embodied in their behavior, e.g., the historical inputs of these users. A good translation should implicitly mirror user traits rather than

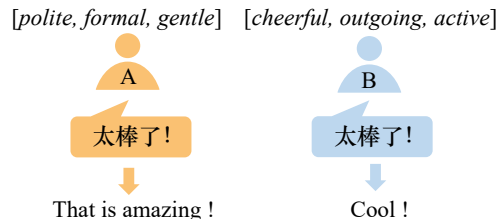


Figure 1: An example in which user traits leads to synonymous yet stylistically different translations.

merely translate the original content, as the example shown in Figure 1. However, current NMT models are mainly designed for the semantic transformation between the source and target sentences regardless of subtle traits with respect to user behavior. It can be said that the effect of user behavior on translation modeling is still far from utilization, which, to some extent, limits the applicability of NMT models in real-world scenarios.

More recently, several studies have shown that the prominent signals in terms of personal characteristics can be served as inductive biases and reflected in translation results using domain adaptation approaches, such as personality (Mirkin et al., 2015), gender (Rabinovich et al., 2017), and politeness (Sennrich et al., 2016a). However, previously explored signals characterize users from a single dimension, which insufficiently represent fine-grained user traits. Furthermore, Michel and Neubig (2018) pay their attention to personalized TED talk translation, in which they train a speaker-specific bias to revise the prediction distribution. In contrast with these studies, our work investigates a more realistic online scenario: a real-world MT system serves extensive users, where the user-behavior annotated data covering all users is unavailable. Previous methods (Mirkin et al., 2015; Michel and Neubig, 2018) require the users in the training set and the test set to be consistent, therefore can not

* Jinsong Su is the corresponding author. This work was done when Huan Lin was interning at DAMO Academy, Alibaba Group.

¹We release our source code and the associated benchmark at <https://github.com/DeepLearnXMU/User-Driven-NMT>.

deal with this zero-shot issue.

Starting from this concern, we explore user-driven NMT that generates personalized translations for users unseen in the training dataset according to their behavior. Specifically, we choose the historical inputs to represent user behavior since they can not only be easily obtained in the real-world scenarios, but also reflect the topic preference, stylistic characteristic, and context of user. Moreover, compared with pre-defined or user-specific labels, historical inputs can be updated with current source sentences, which is also in line with realistic scenario.

In this work, we propose a novel framework for this task, where the NMT model is equipped with a cache module to restore and update historical inputs. Besides, in order to further transfer the traits from the seen users to the unseen ones, we design a regularization framework based on contrastive learning (Bose et al., 2018; Yang et al., 2019), which forces our model to decrease the divergence between translations of similar users while increasing the diversity on dissimilar users.

In order to further train and assess the proposed framework, we construct a new **User-Driven Machine Translation** dataset called **UDT-Corpus**. This corpus consists of 6,550 users with totally 57,639 Chinese sentences collected from a real-world online MT system. Among them, 17,099 Chinese sentences are annotated with their English translations by linguistic experts according to the user-specific historical inputs. Experimental results demonstrate that the proposed framework facilitates the translation quality, and exactly generates diverse translations for different users.

To summarize, major contributions of our work are four-fold:

- We introduce and explore user-driven NMT task that leverages user behavior to enhance translation model. We hope our study can attract more attention to explore techniques on this topic.
- We propose a novel framework for user-driven NMT based on cache module and contrastive learning, which is able to model user traits in zero-shot scenarios.
- We collect UDT-Corpus and make it publicly available, which may contribute to the subsequent researches in the communities of NMT and user-driven models.
- Extensive analyses indicate the effectiveness

of our work and verify that NMT can profit from user behavior to generate diverse translations conforming to user traits.

2 Related Work

This section mainly includes the related studies of personalized machine translation, cache-based NMT and contrastive learning for NMT.

Personalized Machine Translation Recently, some researchers have employed domain adaptation (Zhang et al., 2019; Gururangan et al., 2020; Yao et al., 2020) to generate personalized translations. For example, Mirkin et al. (2015) show that the translation generated by the SMT model has an adverse effect on the prediction of author personalities, demonstrating the necessity of personalized machine translation. Furthermore, Sennrich et al. (2016a) control the politeness in the translation by adding a politeness label on the source side. Rabinovich et al. (2017) explore a gender-personalized SMT system that retains the original gender traits. These domain labels represent users in single dimension separately, which are insufficient to distinguish large-scale users in a fine-grained way. The most correlated work to ours is Michel and Neubig (2018) which introduces a speaker-specific bias into the conventional NMT model. However, these methods are unable to deal with users unseen at the training time. Different from them, user-driven NMT can generate personalized translations for these unseen users in a zero-shot manner.

Cache-Based Machine Translation Inspired by the great success of cache on language modeling (Kuhn and de Mori, 1990; Goodman, 2001; Federico et al., 2008), Nepveu et al. (2004) propose a cache-based adaptive SMT system. Tiedemann (2010) explore a cache-based translation model that fills the cache with bilingual phrase pairs extracted from previous sentence pairs in a document. Bertoldi et al. (2013) use a cache mechanism to achieve online learning in phrase-based SMT. Gong et al. (2011), Kuang et al. (2018), and Tu et al. (2018) further exploit cache-based approaches to leverage contextual information for document-level machine translation. Contrast with the document-level NMT that learns to capture contextual information, our study aims at modeling user traits, such as, topic preference, stylistic characteristics, and expression habits. Moreover, historical inputs of user has relatively fewer dependencies than the contexts

used in document-level translation.

Contrastive Learning for NMT Contrastive learning has been extensively applied in the communities of computer vision and natural language processing due to its effectiveness and generality on self-supervised learning (Vaswani et al., 2013; Mnih and Kavukcuoglu, 2013; Liu and Sun, 2015; Bose et al., 2018). Towards raising the ability of NMT in capturing global dependencies, Wiseman and Rush (2016) first introduce contrastive learning into NMT, where the ground-truth translation and the model output are considered as the positive and contrastive samples, respectively. Yang et al. (2019) construct contrastive examples by deleting words from ground-truth translation to reduce word omission errors in NMT. Contrast to these studies, we employ contrastive learning to create broader learning signals for our user-driven NMT model, where the prediction distribution of translations with respect to similar users and dissimilar users are considered as positive and contrastive samples, respectively. Thus, our model can better transfer the knowledge of the seen users to the unseen ones.

3 User-Driven Translation Dataset

In order to build a user-driven NMT system, we construct a new dataset called UDT-Corpus containing 57,639 inputs of 6,550 users, 17,099 among them are Chinese-to-English translation examples.

3.1 Data Collection and Preprocessing

We collect raw examples from Alibaba Translate² which contain the user inputs and the translations given by the translation system.

For data preprocessing, we first anonymize data and perform data deduplication within each user. Then, we utilize a pre-trained n-gram language model KenLM³ to filter out translation examples with low-quality source data. Moreover, we remove such pairs whose source sentence is shorter than 2 words or longer than 100 words.

3.2 Data Annotation

In the corpus, we represent each translation example as a triplet $\langle X^{(u)}, Y^{(u)}, H^{(u)} \rangle$, where $H^{(u)}$ is the historical inputs of the user u , $X^{(u)}$ is the current source sentence and $Y^{(u)}$ is the target translation sentence annotated with $H^{(u)}$. To obtain

²https://www.aliyun.com/product/ai/base_alimt

³<https://github.com/kpu/kenlm>.

such a triplet, we first sequentially sample up to 10 source sentences which are the historical inputs of each user. Then, for the given historical inputs, we collect their followed source input paired with the pseudo translation given by the translation system. Afterwards, we assign these historical inputs and the current input pairs to two professional annotators and ask them to revise the pseudo translation according to the source sentence and historical inputs. Specifically, we first ask one of them to annotate and the other to evaluate, and then resolve annotation disagreements by reviewing. During annotation, 91.8% of the original data are revised. Moreover, annotators are asked to record whether their revision is affected by user history. The result shows that 76.25% of the sentences are impacted.

4 User-Driven NMT Framework

In this section, we first give a brief description about the problem formulation of user-driven NMT, and then introduce our proposed framework in detail. We choose Transformer (Vaswani et al., 2017) as the basic NMT model due to its competitive performance. In fact, our framework is transparent and applicable to other NMT models.

Figure 2 illustrates the basic framework of the proposed user-driven NMT. Most typically, we equip the NMT model with two user-specific caches to exploit user behavior for better translation (See Section § 4.2). Besides, we augment the conventional NMT training objective with contrastive learning, which allows the model to learn translation diversity across users (See Section § 4.3).

4.1 Problem Formulation

Given the source sentence X and the previously generated words $Y_{<i} = y_1, \dots, y_{i-1}$, the conventional NMT model with parameter θ predicts the current target word y_i by $P(y_i|X, Y_{<i}; \theta)$. As a significant extension of conventional NMT, user-driven NMT with parameter θ aims to model $P(y_i^{(u)}|X^{(u)}, Y_{<i}^{(u)}, u; \theta)$, that is, generates the translation that can reflect the traits of user u . Unlike previous studies (Mirkin et al., 2015; Michel and Neubig, 2018) only caring for generating translations for users seen at the training time, our user-driven NMT mainly focuses on a more realistic online MT scenario, where the users for testing are unseen in the training dataset. Moreover, the conventional domain adaptation methods can not be directly applied to this zero-shot scenario.

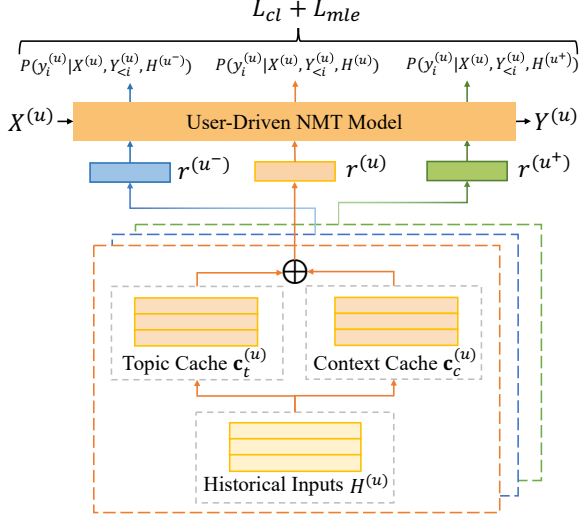


Figure 2: The architecture of our user-driven NMT model. We use the topic cache and context cache to capture the long-term and short-term user traits for user u from corresponding historical inputs $H^{(u)}$, respectively. Then, we combine the representations of two caches to get a user behavior representation $r^{(u)}$, which is fed into the NMT model for personalized translation. Furthermore, we use contrastive learning involving similar user u^+ and dissimilar user u^- to increase the translation diversity among different users.

4.2 Cache-based User Behavior Modeling

Due to the advantages of cache mechanism on dynamic representations (Gong et al., 2011; Kuang et al., 2018; Tu et al., 2018), we equip the conventional Transformer-based NMT model with two user-specific caches to leverage user behavior for NMT: 1) *topic cache* $\mathbf{c}_t^{(u)}$ that aims at capturing the global and long-term traits of user u ; and 2) *context cache* $\mathbf{c}_c^{(u)}$, which is introduced to capture the short-term traits from the recent source inputs of user u . During this process, we focus on the following three operations on cache:

Cache Representation In order to facilitate the efficient computation of the user behavior encoded by our caches, we define each cache as an embedding sequence of keywords. We first calculate TF-IDF values of input words, and then extract words with TF-IDF weights higher than a predefined threshold to represent user behavior.

Note that the calculation of TF-IDF value of a word mainly depends on its frequency in the *document* and inverse document frequency in the *corpus*. Since two caches play different roles in the user-driven NMT model, we identify keywords for two caches based on different definitions of “document”

and “corpus”. Specifically, when constructing topic cache $\mathbf{c}_t^{(u)}$, we treat the historical inputs $H^{(u)}$ of the user u as the “document” and the historical inputs $H^{(u)}$ of all users U as the “corpus”, then define topic cache $\mathbf{c}_t^{(u)}$ as an embedding sequence of historical keywords. Unlike the topic cache, for context cache $\mathbf{c}_c^{(u)}$, we individually consider the current source sentence $X^{(u)}$ and historical inputs $H^{(u)}$ as the TF-IDF “document” and “corpus”, defining $\mathbf{c}_c^{(u)}$ as an embedding sequence of current keywords.

Besides, in the real-world MT scenario, there exists a large number of users without any historical input. For these users, we find the most similar user according to the cosine similarity based on their TF-IDF bag-of-words representations of topic keywords, and initialize the corresponding topic cache with that of the most similar user.

Updating Caches When using an online MT system, users often continuously input multiple sentences. Thus, our caches should be dynamically updated to ensure the accurate encoding of user behavior.

To update topic cache, we first recalculate the TF-IDF values of all historical input words, so as to redetermine the keywords stored in this cache. As for context cache, we consider it as a *filter window* sliding across historical inputs, and apply *first-in-first-out* rule to replace its earliest keywords with the recently input ones.

Reading from Caches During the translation of the NMT model, we perform a gating operation on $\mathbf{c}_t^{(u)}$ and $\mathbf{c}_c^{(u)}$, producing a vector $r^{(u)}$ that reflects user behavior as follows:

$$r^{(u)} = \alpha \bar{\mathbf{c}}_t^{(u)} + (1 - \alpha) \bar{\mathbf{c}}_c^{(u)} \quad (1)$$

$$\alpha = \text{Sigmoid}(\mathbf{W}_t \bar{\mathbf{c}}_t^{(u)} + \mathbf{W}_r \bar{\mathbf{c}}_c^{(u)}), \quad (2)$$

$$\bar{\mathbf{c}}_t^{(u)} = \text{MeanPooling} [\mathbf{c}_t^{(u)}], \quad (3)$$

$$\bar{\mathbf{c}}_c^{(u)} = \text{MeanPooling} [\mathbf{c}_c^{(u)}], \quad (4)$$

where both W_t and W_r are learnable parameter matrices. Then, we directly add $r^{(u)}$ into the embedding sequence of original current source sentence $X^{(u)}$, forming a source embedding sequence with user behavior as follows:

$$\hat{X}^{(u)} = \{x_i^{(u)} + r^{(u)}\}_{1 \leq i \leq |X^{(u)}|}. \quad (5)$$

Finally, the NMT model is fed with $\hat{X}^{(u)}$ to generate the translation for u . Due to the limitation

of pages, we omit the detailed descriptions of the NMT model. Please refer to Vaswani et al. (2017) for the details.

4.3 Model Training with a Contrastive Loss

Given training instances $\langle X^{(u)}, Y^{(u)}, H^{(u)} \rangle$, we train the user-driven NMT model using the following objective function:

$$\mathcal{L} = \mathcal{L}_{\text{mle}} + \mathcal{L}_{\text{cl}}. \quad (6)$$

Here, \mathcal{L}_{mle} is the maximum likelihood translation loss extended from the conventional NMT training objective. Formally, it is defined as:

$$\mathcal{L}_{\text{mle}} = \sum_i -\log P(y_i^{(u)} | X^{(u)}, Y_{<i}^{(u)}, H^{(u)}; \theta). \quad (7)$$

\mathcal{L}_{cl} is a triplet-margin-based contrastive loss, which allows the NMT model to learn the translation diversity across users.

Specifically, for an input sentence, an ideal user-driven NMT model should be able to generate translations with non-divergent user traits for similar users, while producing translations with diverse user traits for dissimilar users. However, using only \mathcal{L}_{mle} cannot guarantee this since it separately considers each training instance during the model training. To deal with this issue, for each training instance $\langle X^{(u)}, Y^{(u)}, H^{(u)} \rangle$, we first determine the most similar user u^+ according to the cosine similarity based on their bag-of-keyword representations, and randomly select a user without any same keyword as the dissimilar user u^- of u . Finally, using historical inputs of u^+ and u^- , we construct several pseudo training instances to define \mathcal{L}_{cl} as follows:

$$\mathcal{L}_{\text{cl}} = \sum_{u \in U} \max[d(X^{(u)}, Y^{(u)}, H^{(u)}, H^{(u^+)}) - d(X^{(u)}, Y^{(u)}, H^{(u)}, H^{(u^-)}) + \eta, 0], \quad (8)$$

where $d(X^{(u)}, Y^{(u)}, H^{(u)}, H^{(u^+)})$

$$= \left\| \frac{1}{|Y^{(u)}|} \sum_i \log P(y_i^{(u)} | X^{(u)}, Y_{<i}^{(u)}, H^{(u)}) - \frac{1}{|Y^{(u)}|} \sum_i \log P(y_i^{(u)} | X^{(u)}, Y_{<i}^{(u)}, H^{(u^+)}) \right\|^2 \quad (9)$$

and η is a predefined threshold, which is set to 2 in our experiments. Here, we omit the definition of

| | Train | Dev | Test |
|-------------------------|--------|-------|-------|
| #user | 5,350 | 600 | 600 |
| #historical input | 33,441 | 3,629 | 3,470 |
| #current sentence pairs | 14,006 | 1,557 | 1,536 |

Table 1: Dataset for fine-tuning experiments.

$d(X^{(u)}, Y^{(u)}, H^{(u)}, H^{(u^-)})$, which is similar to $d(X^{(u)}, Y^{(u)}, H^{(u)}, H^{(u^+)})$.

Formally, \mathcal{L}_{cl} will encourage the NMT model to minimize the prediction difference between the training instances $\langle X^{(u)}, Y^{(u)}, H^{(u)} \rangle$ and $\langle X^{(u)}, Y^{(u)}, H^{(u^+)} \rangle$, and maximize the difference between the training instances $\langle X^{(u)}, Y^{(u)}, H^{(u)} \rangle$ and $\langle X^{(u)}, Y^{(u)}, H^{(u^-)} \rangle$. In this way, the NMT model can not only exploit pseudo training instances, but also produce more consistent translations with user traits.

5 Experiments

In this section, we carry out several groups of experiments to investigate the effectiveness of our proposed framework on UDT-Corpus.

5.1 Setup

We develop the user-driven NMT model based on Open-NMT Transformer (Klein et al., 2017), and adopt a two-stage strategy to train this model: we first pre-train a Transformer-based NMT model on the WMT2017 Chinese-to-English dataset, and then fine-tune this model to our user-driven NMT model using UDT-Corpus.

Datasets The WMT2017 Chinese-to-English dataset is composed of the News Commentary v12, UN Parallel Corpus v1.0, and CWMT corpora, with totally 25M parallel sentences. To fine-tune our model, we split UDT-Corpus into training, validation and test set, respectively. Table 1 provides more detailed statistics of these datasets. To improve the efficiency of model training, we train the model using only parallel sentences with no more than 100 words. Following common practices, we employ byte pair encoding (Sennrich et al., 2016b) with 32K merge operations to deal with all sentences.

Training Details Following Vaswani et al. (2017), we use the following hyper-parameters: the word embedding dimension is set to 512, the hidden layer dimension is 2048, the layer numbers of

both encoder and decoder are set to 6, and the number of attention heads is set to 8. Besides, we use 4 GPUs for training. At the pre-training stage, we employ the Adam optimizer with $\beta_2 = 0.998$. We use the batch size of 16,384 tokens and pre-train the model for 200,000 steps. Particularly, we adopt the dropout strategy (Srivastava et al., 2014) with rate 0.1 to enhance the robustness of our model. When fine-tuning the model, we keep the other settings consistent with the pre-training stage, but reduce the batch size to 2048 tokens and fine-tune the model with early-stopping strategy.

Evaluation We assess the translation quality with two metrics: one is case-insensitive BLEU (*mteval-v13a.pl*, Papineni et al., 2002)⁴ and the other is METEOR⁵ (Denkowski and Lavie, 2011).

5.2 Baselines

We represent our user-driven NMT model as **UD-NMT** and compare it with the following baselines:

- **TF**. It is a Transformer-based NMT model pre-trained on the WMT2017 corpus. This model yields 24.61 BLEU score on WMT2017 Chinese-to-English translation task, which is comparable with reported results in (Wan et al., 2020; Zhou et al., 2020), which makes our subsequent experiments convincing.
- **TF-FT**. This model is also a Transformer-based NMT model that is further fine-tuned on the parallel sentences of UDT-Corpus.
- **TF-FT + PesuData**. This model is a variant of *TF-FT*. When constructing it, we pair historical inputs with their translations produced by our online translation system, forming additional data for fine-tuning *TF-FT*.
- **TF-FT + ConcHist** (Tiedemann and Scherrer, 2017). In this model, we introduce user behavior into *TF-FT* by concatenating each input sentence with several historical inputs. We mark all tokens in historical inputs with a special prefix to indicate that they are additional information.
- **TF-FT + UserBias** (Michel and Neubig, 2018). It introduces user-specific biases to refine softmax-based predictions of Transformer NMT model. We change it to a zero-shot method similar to (Farajian et al., 2017)

⁴<https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl>

⁵<https://github.com/cmu-mtlab/meteor>

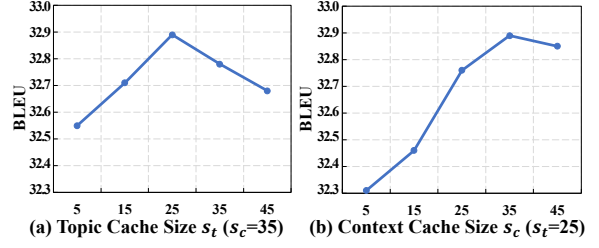


Figure 3: Effects of cache size on translation quality.

| Model | BLEU | METEOR |
|--------------------------|--------------|--------------|
| <i>w/o user behavior</i> | | |
| TF | 27.52 | 44.05 |
| TF-FT | 28.61 | 45.35 |
| TF-FT + PesuData | 29.02 | 45.40 |
| <i>w/ user behavior</i> | | |
| TF-FT + ConcHist | 30.85 | 46.08 |
| TF-FT + UserBias | 31.36 | 46.79 |
| UD-NMT | 32.35 | 48.20 |

Table 2: Main results on UDT-Corpus. “w/o”, “w/” denote “without” and “with”, respectively.

since (Michel and Neubig, 2018) can not be directly applied to our scenario. In particular, we replace the user ID in the test set with that of the most similar user in the training set.

Note that the first two baselines, e.g., *TF* and *TF-FT*, are conventional NMT models without exploiting user behavior.

5.3 Effect of Cache Sizes

Since cache size directly determines the utility of user behavior, we investigate its effect on the performance of **UD-NMT**. We denote the sizes of topic cache and context cache as s_t and s_c for simplicity.

Figure 3 lists the performance of our model with different s_t and s_c on validation set. We observe that s_t larger than 25 and s_c larger than 35 do not lead to significant improvements. For this result, we speculate that small cache sizes are unable to capture sufficient user behavior for NMT. However, since the number of keywords are limited, larger cache sizes only bring limited information gain. Therefore, we directly use $s_t = 25$ and $s_c = 35$ in the subsequent experiments.

5.4 Main Results

From Table 2, we observe that our *UD-NMT* model consistently outperforms all baselines in terms of two metrics. Moreover, we draw several interesting conclusions:

| Model | BLEU \uparrow | METEOR \uparrow | s-BLEU \uparrow | d-BLEU \uparrow | s-Sim. \downarrow | d-Sim. \downarrow |
|---------------------------------|-----------------|-------------------|-------------------|-------------------|---------------------|---------------------|
| UD-NMT | 32.35 | 48.20 | 32.17 | 32.23 | 93.18 | 80.10 |
| w/o topic cache | 31.88 \dagger | 48.00 | — | — | — | — |
| w/o context cache | 31.86 \dagger | 47.84 \dagger | 31.94 \dagger | 31.58 \dagger | 88.61 | 69.32 |
| w/o similar user initialization | 32.02 | 48.14 | 31.86 \dagger | 31.13 \ddagger | 93.54 \dagger | 80.16 |
| w/o contrastive learning | 32.00 | 48.09 | 31.88 \dagger | 31.94 | 93.49 \dagger | 81.59 \dagger |

Table 3: Ablation Study. \uparrow : higher is better, \downarrow : lower is better. Since the user similarity is calculated based on the topic keywords, the model can not find similar user and dissimilar user without it. Thus *w/o topic cache* does not have the s-BLEU, s-Sim., d-BLEU and d-Sim.. \ddagger/\dagger : indicates the drop of translation quality is statistically significant comparing to “UD-NMT” ($p < 0.01/0.05$).

1) All NMT models leveraging user behavior surpass vanilla models, including *TF*, *TF-FT*, showing that user behavior is useful for NMT.

2) *UD-NMT* exhibits better than *TF-FT + Pesu-Data*, which uses the same training data as ours. The underlying reason is that *UD-NMT* can leverage user traits to generate better translations.

3) Although both *TF-FT + UserBias* and *UD-NMT* exploit user behavior for NMT, *UD-NMT* achieves better performance than *TF-FT + UserBias* without introducing extra parameters. This result demonstrates the advantage of cache on modeling user behavior than introducing user-specific biases into model parameters.

5.5 Ablation Study

To explore the effectiveness of different components in our model, we further compare *UD-NMT* with its several variants, as shown in Table 3.

Particularly, we propose to evaluate translations using the following variant metrics: **s-BLEU**, **s-Sim.**, **d-BLEU** and **d-Sim.**. When using s-BLEU, we replace the topic cache of current user with that of his most similar user. Keeping the same current input, we calculate the BLEU score with ground-truth as reference and the translation for this similar user as hypothesis. As for s-Sim., we adopt the same strategy as s-BLEU, but use the translation for original user as reference to evaluate the BLEU score. In other words, s-BLEU and d-BLEU assesses the translation quality given unsuitable user. Therefore, higher s-BLEU and d-BLEU indicates better model robustness, while s-BLEU and d-BLEU measures how much the translation changes given different user. Thus lower s-Sim. and d-Sim. show larger translation diversity.

Our conclusions are shown as follows:

1) *w/o topic cache*. To build this variant, we remove topic cache from our model. The result in Line 2 indicates that removing topic cache leads to

a performance drop, suggesting that topic cache is useful for modeling user behavior.

2) *w/o context cache*. Unlike the above variant, we only use topic cache to represent user traits in this variant. According to the results shown in Line 3, we observe that this change results in a significant performance decline of our model, demonstrating that context cache also effectively captures user behavior for NMT. However, the translation diversity among users increases since the model will not be affected by the context cache in this variant, which is the same between different users when calculating s-Sim. and d-Sim..

3) *w/o similar user initialization*. In this variant, we do not initialize topic caches of the users without historical inputs using that of the most similar users. From Line 4, we observe that the performance of our model degrades without similar user initialization.

4) *w/o contrastive learning*. In this variant, we remove the contrastive learning from the whole training objective to inspect the performance change of our model. As shown in Line 4, the performance of our model drops, proving that the contrastive learning is important for the training of our model.

Moreover, we can infer from **Column 6** and **7** that our model can generate diverse translations. Specifically, the translations of dissimilar users has larger diversity than that of similar ones. Furthermore, we conclude that our model is robust, since it still performs well when we replace the topic cache of current user with those of other users (See **Column 4** and **5**).

5.6 Analysis of Contrastive Margin

Inspired by Yang et al. (2019), we argue that the contrastive learning may increase the prediction diversity of our model between users compared with using the MLE loss. To confirm this, we randomly

| | | |
|---------------|--------------------|---|
| Src | | 基因芯片分析发现, 在 crf 多突变体 中, 许多受b型arrs 调控的基因同时也受crfs的调制剂起负反馈调节作用。 |
| Ref | | Gene chip analysis found out that in multiple CRF mutants, many genes regulated by type b arrs were also negatively regulated by the modulators of CRFs. |
| User A | Topic Cache | 木马 (trojan) 渔夫 (fisherman) 百洁布 (cleaning cloth) 耳机 (earphone) 蓝牙 (bluetooth) |
| | Translation | Gene chip analysis found that in the CRF mutant, many genes regulated by b arrs are also negatively fed by CRFs toning . |
| User B | Topic Cache | 乙烯 (ethylene) 伸长 (extension) 促进 (promote) 基因 (gene) 生长素 (auxin) 调控 (regulate) |
| | Translation | Gene chip analysis found that in the CRF mutant, many genes regulated by type b arrs are also subject to negative feedback adjustment by CRFs modulating agents. |

(a) Translations of UD-NMT given different user traits. Words highlighted in green indicate the unnatural or awkward translations.

| | | |
|--------------------------|--|---|
| Historical inputs | | 面料成分 氨纶 风格 性感 款式 连体裤 颜色 白色, 黑色 (Fabric Composition Spandex Style Sexy Type Jumpsuit Color White, Black) 2020 秋冬季新款港风复古落肩外套女宽松学生毛绒短款上衣 (2020 Autumn and Winter New Hong Kong Style Retro Drop Sleeves Jacket Female Loose Student Plush Crop Top) |
| Src | | 牛津纺面料防水 耐磨 , 15英寸大小 |
| Ref | | Oxford Fabric Waterproof and Wear Resistant , 15 Inch in Size |
| TF-FT + PesuData | | Oxford Textile Fabrics Waterproof and Waterproof , 15 Inches Size |
| TF-FT + UserBias | | Oxford Woven Fabric is Waterproof and Resistant , 15 Inches in Size |
| UD-FT | | Oxford Woven Fabric Waterproof and Wear - Resistant , 15 Inch Size |

(b) Translations of different models. Words highlighted in red are incorrect translations.

Figure 4: Two examples of user-driven machine translation.

sample 300 examples from the training dataset, and compute the following margin:

$$\Delta = \left[d^{(u^+)}(\cdot) - d^{(u^-)}(\cdot) \right] - \left[d_{\text{mle}}^{(u^+)}(\cdot) - d_{\text{mle}}^{(u^-)}(\cdot) \right],$$

where $d^{(u^+)}(\cdot)$ is defined in Equation 9. The definition of $d_{\text{mle}}^{(u^+)}(\cdot)$ is the same with $d(\cdot)$, the only difference lies in that the NMT model is only trained by the conventional MLE loss. We find that $d(\cdot)$ has a larger margin than $d_{\text{mle}}(\cdot)$ on 88% of sampled sentence pairs, with an average margin of 0.19. The results indicate again that the contrastive learning increases the translation diversity.

5.7 Qualitative Analysis

In order to intuitively understand how our cache module exactly affects the translations, we feed our model with the same current source sentence but different users, and display the 1-best translations generated by our model. As shown in the Figure 4 (a), our model is able to produce correct but diverse translations according to different topic caches. Moreover, it is interesting to observe that specific topic keywords such as “type b arr”, “negatively regulated” and “modulators” are translated to synonymous but “out-of-domain” phrases if the topic cache does not conform to input sentence. On the contrary, the model conversely generates “in-domain” translation if the topic cache comes from the same topic of input sentence.

| Correlation Order | Proportion |
|---------------------------|------------|
| UD-NMT > TF-FT + PesuData | 86% |
| UD-NMT > TF-FT + UserBias | 74% |

Table 4: The proportion of translations more related to historical inputs assessed by human translators. A > B indicates the translations generated by A system is more correlated to history inputs.

Besides, to further reveal the effect of user behavior, we provide an example in Figure 4 (b), which lists different translations by compared models for the same inputs. The historical inputs indicate that this user may be an apparel seller, since his historical inputs contain the product titles and descriptions of clothing. Thus, the keywords “Wear Resistant” in the source sentence are correlated with this user. However, two baselines translate it to “Waterproof” and “Resistant”, respectively. Moreover, *TF-FT + UserBias* generates a subject-verb-object structured sentence by adding the auxiliary verb “is”, which does not conform to the expression habit of the product title. By contrast, with the hint of the keywords in historical inputs, our *UD-NMT* is able to produce suitable translation consistent with the topic preference of this user.

5.8 Manual Evaluation

To further find out whether the improvements of our model are contributed by user traits, we ran-

domly sample 100 examples from the test dataset and ask the linguist experts to sort different systems according to the relevance between the generated translations and the historical input. The results in Table 4 show that our model can generate translations more in line with history inputs than baseline models in most cases, proving that our method can make better use of user traits.

6 Conclusion

We propose user-driven NMT task, which aims to leverage user behavior to generate personalized translations. With the help of cache module and contrastive estimation, we successfully build an end-to-end NMT model that is able to capture potential user traits from their historical inputs and generate diverse translations under a zero-shot learning fashion. Furthermore, we contribute UDT-Corpus, which is the first Chinese-English parallel corpus annotated with user behavior. We expect our study can attract more attention towards this topic. It is a promising direction to explore other behavior in future, such as clickthrough and editing operations. Moreover, following recent advancements in domain adaptation for NMT, we plan to further improve our model via adversarial training based knowledge transfer (Zeng et al., 2018; Yao et al., 2020; Su et al., 2021) and dual knowledge transfer (Zeng et al., 2019).

Acknowledgments

The project was supported by National Key Research and Development Program of China (No. 2020AAA0108004 and No. 2018YFB1403202), National Natural Science Foundation of China (No. 61672440), Natural Science Foundation of Fujian Province of China (No. 2020J06001), Youth Innovation Fund of Xiamen (No. 3502ZZ20206059), and the Fundamental Research Funds for the Central Universities (No. ZK20720200077). We also thank the reviewers for their insightful comments.

References

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. [Cache-based online adaptation for machine translation enhanced computer assisted translation](#). In *Proceedings of the 2013 Machine Translation Summit*, pages 35–42.

Avishek Joey Bose, Huan Ling, and Yanshuai Cao. 2018. [Adversarial contrastive estimation](#). In *Proceedings of the 56th Annual Meeting of the Asso-*

ciation for Computational Linguistics, pages 1021–1032.

- Michael J. Denkowski and Alon Lavie. 2011. [Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems](#). In *Proceedings of the 6th Workshop on Statistical Machine Translation*, pages 85–91.
- M. Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. [Multi-domain neural machine translation through unsupervised adaptation](#). In *Proceedings of the Second Conference on Machine Translation*, pages 127–137.
- Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. [IRSTLM: an open source toolkit for handling large scale language models](#). In *Proceedings of the 9th Annual Conference of the International Speech Communication Association*, pages 1618–1621.
- Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. [Cache-based document-level statistical machine translation](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 909–919.
- Joshua T. Goodman. 2001. [A bit of progress in language modeling](#). *Computer Speech and Language*, 15(4):403–434.
- Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. [Don’t stop pretraining: Adapt language models to domains and tasks](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.
- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senelart, and Alexander Rush. 2017. [OpenNMT: Open-source toolkit for neural machine translation](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 67–72.
- Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2018. [Modeling coherence for neural machine translation with dynamic and topic caches](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 596–606.
- Roland Kuhn and Renato de Mori. 1990. [A cache-based natural language model for speech recognition](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6):570–583.
- Yang Liu and Maosong Sun. 2015. [Contrastive unsupervised word alignment with non-local features](#). In *Proceedings of the 29th Association for the Advancement of Artificial Intelligence*, pages 2295–2301.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. [Effective approaches to attention-based neural machine translation](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

- Paul Michel and Graham Neubig. 2018. [Extreme adaptation for personalized neural machine translation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 312–318.
- Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. [Motivating personality-aware machine translation](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1108.
- Andriy Mnih and Koray Kavukcuoglu. 2013. [Learning word embeddings efficiently with noise-contrastive estimation](#). In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pages 2265–2273.
- Laurent Nepveu, Guy Lapalme, Philippe Langlais, and George F. Foster. 2004. [Adaptive language and translation models for interactive machine translation](#). In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 190–197.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.
- Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. [Personalized machine translation: Preserving original author traits](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1074–1084.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. [Controlling politeness in neural machine translation via side constraints](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 35–40.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. [Dropout: a simple way to prevent neural networks from overfitting](#). *The Journal of Machine Learning Research*, 15(1):1929–1958.
- Jinsong Su, Jiali Zeng, Jun Xie, Huating Wen, Yongjing Yin, and Yang Liu. 2021. [Exploring discriminative word-level domain contexts for multi-domain neural machine translation](#). *IEEE Trans. Pattern Anal. Mach. Intell.*, 43:1530–1545.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. [Sequence to sequence learning with neural networks](#). In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems*, pages 3104–3112.
- Jörg Tiedemann. 2010. [Context adaptation in statistical machine translation using models with exponentially decaying cache](#). In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 8–15.
- Jörg Tiedemann and Yves Scherrer. 2017. [Neural machine translation with extended context](#). In *Proceedings of the 3rd Workshop on Discourse in Machine Translation*, pages 82–92.
- Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. [Learning to remember translation history with a continuous cache](#). *Transactions of the Association for Computational Linguistics*, 6:407–420.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Proceedings of the 31th Annual Conference on Neural Information Processing Systems*, pages 5998–6008.
- Ashish Vaswani, Yingdong Zhao, Victoria Fossum, and David Chiang. 2013. [Decoding with large-scale neural language models improves translation](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1387–1392.
- Yu Wan, Baosong Yang, Derek F. Wong, Yikai Zhou, Lidia S. Chao, Haibo Zhang, and Boxing Chen. 2020. [Self-paced learning for neural machine translation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1074–1080.
- Sam Wiseman and Alexander M. Rush. 2016. [Sequence-to-sequence learning as beam-search optimization](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306.
- Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019. [Reducing word omission errors in neural machine translation: A contrastive learning approach](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 6191–6196.
- Liang Yao, Baosong Yang, Haibo Zhang, Boxing Chen, and Weihua Luo. 2020. [Domain transfer based data augmentation for neural query translation](#). In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4521–4533. International Committee on Computational Linguistics.

- Jiali Zeng, Yang Liu, Jinsong Su, Yubin Ge, Yaojie Lu, Yongjing Yin, and Jiebo Luo. 2019. [Iterative dual domain adaptation for neural machine translation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 845–855.
- Jiali Zeng, Jinsong Su, Huating Wen, Yang Liu, Jun Xie, Yongjing Yin, and Jianqiang Zhao. 2018. [Multi-domain neural machine translation with word-level domain context discrimination](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 447–457.
- Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. [Curriculum learning for domain adaptation in neural machine translation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1903–1915.
- Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao. 2020. [Uncertainty-aware curriculum learning for neural machine translation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6934–6944.