Front. Comput. Sci., 2022, 16(3): 163344
https://doi.org/10.1007/s11704-022-1297-0

LETTER

Linguistic feature template integration for Chinese-Vietnamese neural machine translation

Zhiqiang YU[1,2,3], Yantuan XIAN[1,3], Zhengtao YU (✉)[1,3], Yuxin HUANG[1,3], Junjun GUO[1,3]

1 Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China
2 School of Mathematics and Computer Science, Yunnan Minzu University, Kunming 650500, China
3 Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, China

1 Introduction and main contributions

Template-based approaches have achieved significant progress in low-resource neural machine translation (NMT) recently [1], such as the efficient works, NMT-GTM [2], SoftPrototype [3], etc. However, most previous works only retrieve target sentence as template to generate translation, neglecting the utilization of linguistic feature that contained in the source sentence and template.

In this paper, we propose a novel NMT approach, LFTI to represent linguistic feature in Chinese and Vietnamese as template and integrate it into translation procedure. In particular, we first investigate a representative linguistic feature named modifier reverse in Chinese and Vietnamese and formalized represent it as linguistic feature template. Apart from this, we use modified sequence-to-sequence NMT architecture to integrate the linguistic feature template. The proposed approach can represent the modifier reverse as linguistic feature template and significantly outperforms the baseline models on Chinese-Vietnamese translation task by integrating the template. Experimental results show that our approach significantly outperforms the baseline models on Chinese-Vietnamese translation task and demonstrates the effectiveness of the linguistic feature template. The contributions of the paper are summarized as follows.

- We investigate a representative linguistic feature named modifier inverse in Chinese and Vietnamese in the perspective of linguistic difference and formalized represent it as modifier sequence.
- We propose a novel translation framework which can integrate the modifier inverse into conventional temp late-based translation architecture.

2 Design and implement of LFTI

To better understand the proposed approach LFTI, we present the necessary background and main technologies used in our framework in this section.

Received June 1, 2021; accepted January 26, 2022

E-mail: ztyu@hotmail.com

2.1 Linguistic feature representation

There are obvious linguistic differences in Chinese-Vietnamese language pair in terms of language features. As the example shown in Fig. 1, for the modification of the center word *candy*, the attributive that modifies the center word in Chinese is always prepositive, while the attributive in Vietnamese is usually postpositive. In Chinese, the basic descriptive attributives are: 1. predicate phrase; 2. verb (phrase)/ preposition phrase; 3. adjective phrases and other descriptive phrases; 4. adjectives and descriptive nouns that without "的". In conventionality condition, the order of descriptive attributive structures in Chinese is 1-2-3-4-center word, while for Vietnamese the order is center word-4-3-2-1. The order of descriptive attributive structures in Vietnamese and Chinese is reversed, it is a very important factor in translating Chinese to Vietnamese and vice versa [4−5].

We summarized this significant language difference feature between Chinese and Vietnamese as modifier reverse. Table A1 in the Online Resource illustrates the three kinds of main modifier sequences that represent the modifier reverse: noun, verb and adjective. Take the noun modifier sequence as an example, the modifier order in Chinese is [adverb]-adjective-noun, while in Vietnamese the order of modifiers is noun-adjective-[adverb].

For integrating modifier reverse into NMT model, we describe it by part of speech (POS) tags. Specifically, standard POS tagging sets, ICTCLAS and VLSP, are chosen to

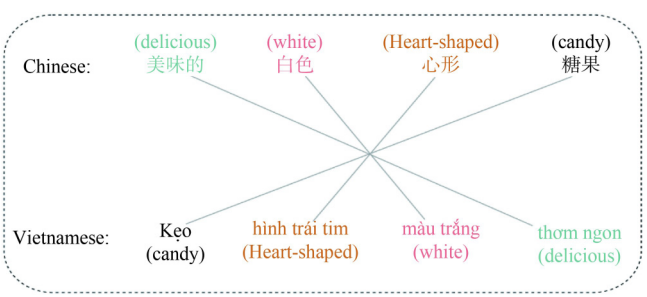

Fig. 1 Modifier reverse in Chinese and Vietnamese

represent Chinese and Vietnamese tagging respectively. Table A2 and Table A3 in the Online Resource list the main POS tags. We retrieve specific noun, verb and adjective modifier sequences by linguistic rule matching. Take the Chinese noun modifier sequence retrieval shown in Algorithm 1 as an example, given the input sentence $x = (x_{1,2,\ldots,m})$ with length m, POS tagging and preprocessing operations are conducted on it, generating POS tag sequence tx. Subsequently, tx is fed into the noun regularized matching block. When the POS tag of tx at time i is n, we judge whether n is preceded by ([d], [a], a), any sequence that does not conform to the linguistic rules will not be recorded. Chinese verb and adjective modifier sequence retrievals are done in the same manner as Algorithm 1. Actually, Vietnamese modifier retrieval is analogous to Chinese modifier retrieval, the slight differences are the specific linguistic rules and the representation of returned tags [6].

Algorithm 1 Modifier sequence retrieval

input: segmented sentence $x = (x_{1,\ldots,m})$ with length m
$tx \leftarrow$ POS(x), get tag sequence by POS tagging
$tx \leftarrow$ preprocess the tag sequence (e.g., convert noun tags $\in \{ng, \ldots, r\}$ to n, convert $a + uj$ tag to a, etc.)
for $i \leftarrow 1$ **to** m **do**
  if $tx_i = n$
    for $k \leftarrow 1$ **to** $i - 1$ **do**
      if $tx_{i-k} = a \vee (tx_{i-k} = d \wedge tx_{i-k+1} = a)$
        pass
      else if $tx_{i-k} \notin \{a, d\} \vee (tx_{i-k} = a \wedge tx_{i-k+1} = d)$
        clean tags $tx_{<i-k}$ and $tx_{>i}$
return tx

2.2 Modified NMT architecture

For integrating the modifier reverse, we propose a modified NMT architecture, by which the modifier reverse feature could be integrated into the NMT inputs. As shown in Fig. 2, we use two encoders to encode source sentence and template into hidden states respectively. To facilitate the integration of the modifier sequence, we infuse the modifier sequence to the original input in embedding layer. For example, if the tags for x_2, x_3 and x_4 are t_2, t_3 and t_4, their representations will be added separately, other untagged tokens are not affected.

3 Evaluation

Datasets We evaluate our approach on the Chinese-Vietnamese translation task, by training models on the ALT dataset. The preprocessed ALT dataset comprises 18,088, 1,000, 1,018 sentence pairs as training, development and test sets respectively.

Baseline We compare our approach with the SMT basic model Moses, NMT basic model Transformer and the homogeneous previous work SoftPrototype [3].

Result The recognition rates are reported in Table 1. In which "noun", "verb" and "adjective" columns denote the recognition rates of the three kinds of modifier sequence respectively, computed as the percentage of times the modifier sequences were recognized out of the original sentence pairs. We observe that the rate of verb modifier sequence is clearly higher than the other two types of modifier sequence, this is consistent with customary human expression. Table 2 shows the experimental results evaluated by BLUE score. Our approach scores between 10.67 and 10.94 BLEU points, which are higher than the best baseline system on zh→vi translation task, there is a similar increasing trend on vi→zh task. Moreover, we note that all types of proposed template strategies (noun, verb, and adjective) indicate the accuracy gains, in which verb template strategy performs better than the other two types.

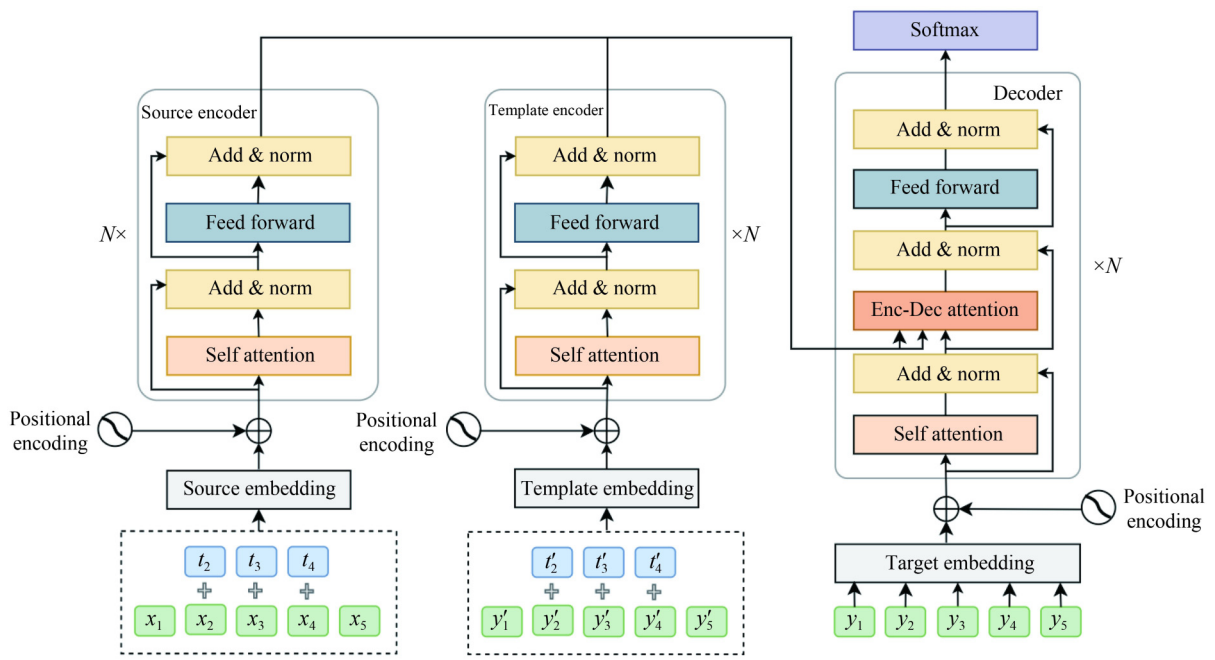

Fig. 2 Overview of our NMT framework. (t_2, t_3, t_4) and (t'_2, t'_3, t'_4) are corresponding modifier sequence for input sentence x and input template y'

Table 1 Recognition rates on ALT datasets

Dataset	Noun/%	Verb/%	Adjective/%
ALT	5.2	7.3	4.9

Table 2 Translation quality evaluation.noun, verb and adjective denote that noun, verb and adjective linguistic feature templates are adopted respectively

Models	ALT	
	zh→vi	vi→zh
Moses	8.84	8.72
Transformer	8.51	8.44
SoftPrototype	9.72	9.37
Our approach (noun)	10.67	10.08
Our approach (verb)	**10.94**	**10.35**
Our approach (adjective)	10.78	10.23

Acknowledgements This work was supported by the National Key Research and Development Plan Project (2019QY1800), and the National Natural Science Foundation of China (Grant Nos. 61732005, 61672271, 61761026, and 61866020)

Supporting information The supporting information is available online at journal.hep.cn and link.springer.com.

References

1. Sennrich R, Zhang B. Revisiting low-resource neural machine translation: a case study. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019, 211–221
2. Cao Q, Xiong D. Encoding gated translation memory into neural machine translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018, 3042–3047
3. Wang Y, Xia Y, Tian F, Gao F, Qin T, Zhai C, Liu T Y. Neural machine translation with soft prototype. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019, 567
4. Tran H A, Huang H, Tran P, Shi S, Nguyen H. Preordering for Chinese-Vietnamese statistical machine translation. IEICE Transactions on Information and Systems, 2019, 102(2): 375–382
5. He J, Yu Z, Lv C, Lai H, Gao S, Zhang Y. Language post positioned characteristic based Chinese-Vietnamese statistical machine translation method. In: Proceedings of the 2017 International Conference on Asian Language Processing (IALP). 2017, 180–184
6. Nguyen D Q, Vu T, Nguyen D Q, Dras M, Johnson M. From word segmentation to POS tagging for Vietnamese. In: Proceedings of the Australasian Language Technology Association Workshop 2017. 2017, 108–113