07

Generating diverse translation with perturbed kNN-MT

# Choose the best translation from diverse candidates


## Abstract

If machine translation can output a variety of translation candidates, we can easily select the appropriate translation from them according to context or TPO. However, the previous method to generate translation candidates only outputs similar candidates, which cannot help users. To generate accurate and diverse translation candidates, we propose a method to generate candidates by searching a wide range of examples from large-scale data. Our method includes a feature of randomly selecting data from retrieved examples. Thus, our approach probabilistically considers a range of examples, leading to generating accurate and diverse translation candidates. In the future, our method will make it easy to edit translation errors by choosing the correct translation from candidates. It might also be possible to adapt the model to the specific domain by changing searched data for more accurate translation.


## Diverse Translation Candidates

このカメは主に何を食べますか？

- What does this turtle mainly eat?
- What does this tortoise mainly eat?


The correct translation may be different based on the context or situation.
→ If the machine translation model can generate several translation candidates, we can easily select the appropriate translation from them.

### Difficulties of generating candidates

In previous methods:

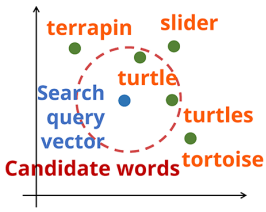
1. Tend to generate similar candidates
2. Candidates are less accurate

### Our research

Our model can generate high-quality and diverse translation candidates by improving the search-based machine translation model.

## Generate Diverse Candidates with Search-based Method

### Previous search-based translation method



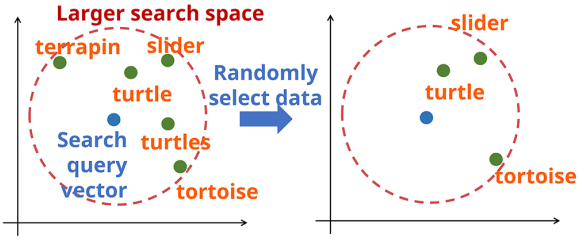

Search output words, which are similar to the input, from a database and use them for machine translation
→ It is known to improve performance.

What does this turtle mainly eat?
What do these turtles mainly eat?

The previous method tends to retrieve similar words from the DB.
Thus, generated candidates are also similar.

U. Khandelwa et al., "Nearest Neighbor Machine Translation," in *Proc. International Conference on Learning Representations* (ICLR), 2021.

### Proposed method for generating diverse candidates [1]




Expand the search space and randomly select the data from large search results.
→ We can use various data and generate both high-quality and diverse candidates.

What does this turtle mainly eat?
What does this tortoise mainly eat?
What does this slider mainly eat?

Using larger search space, we could generate accurate and diverse translation candidates.

## Effect of Our Proposed Method

| | Diversity (DP) ↑ | Translation Accuracy (BLEU) ↑ |
|---|---|---|
| ① Conventional model | 31.4 | 42.6 |
| ② Previous diverse translation method | 35.9 | 40.0 |
| ③ Previous search-based method | 32.3 | 51.8 |
| Proposed A. Combination of ② and ③ | 42.0 | 48.6 |
| Proposed B. ② + ③ + Random selection | 54.4 | 48.4 |

We expand the search space 1.1 to 1.4 times larger for the random selection method.
DP: A metric of how many different words/phrases are included in the output candidate sentences

### Experimental Results

- Previous search-based method (③) used a database for translation and achieved better translation accuracy.
- However, these previous method (①、②、③) tends to generate similar candidates, resulting in smaller DP scores.
- Both our proposed methods, A and B, achieved high DP scores while maintaining translation accuracy.

→ Especially our proposed method B, which randomly selects data from a large search space, benefits from high translation accuracy by a search-based method and a high DP score by a random selection method.


## References

[1] Yuto Nishida, Makoto Morishita, Hidetaka Kamigaito, Taro Watanabe, "Generating Diverse Translation with Perturbed kNN-MT," in *Proc. The 29th Annual Meeting of ANLP* (in Japanese), 2023.



## Contact

Makoto Morishita
Linguistic Intelligence Research Group, Innovative Communication Laboratory