EACL 2024

kNN-BOX: A Unified Framework for Nearest Neighbor Generation

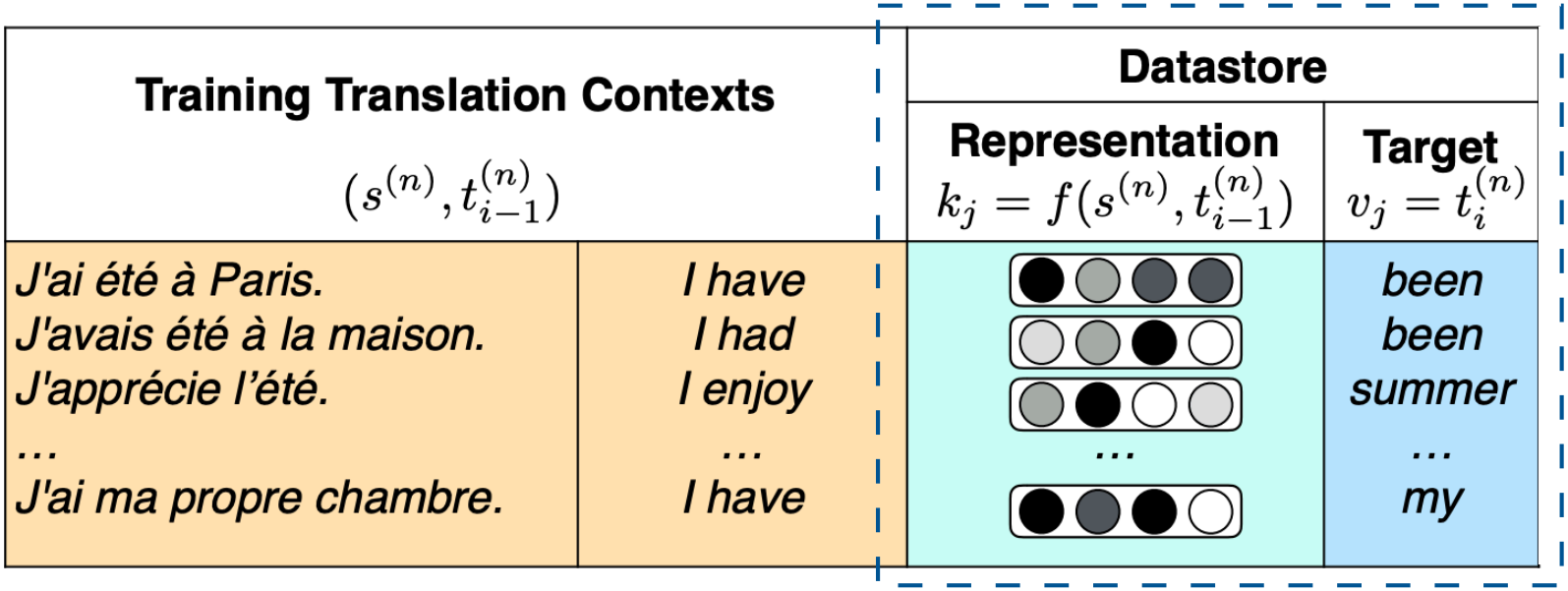
Wenhao Zhu*, Qianfeng Zhao*, Yunzhe Lv*,
Shujian Huang, Siheng Zhao, Sizhe Liu, Jiajun Chen

Nanjing University

* Equal Contributions

Background

- The k-nearest neighbor machine translation (kNN-MT) system incorporates a symbolic datastore to assist NMT model.
- The added symbolic datastore usually saves a huge amount of token-level translation knowledge.



stored knowledge:
generate the value token
at the hidden state key

Figure from: Khandelwal et al. Nearest Neighbor Machine Translation. ICLR'2021.

Background

- During inference, the NMT model will retrieve relevant knowledge from the datastore and use it to refine its original prediction.

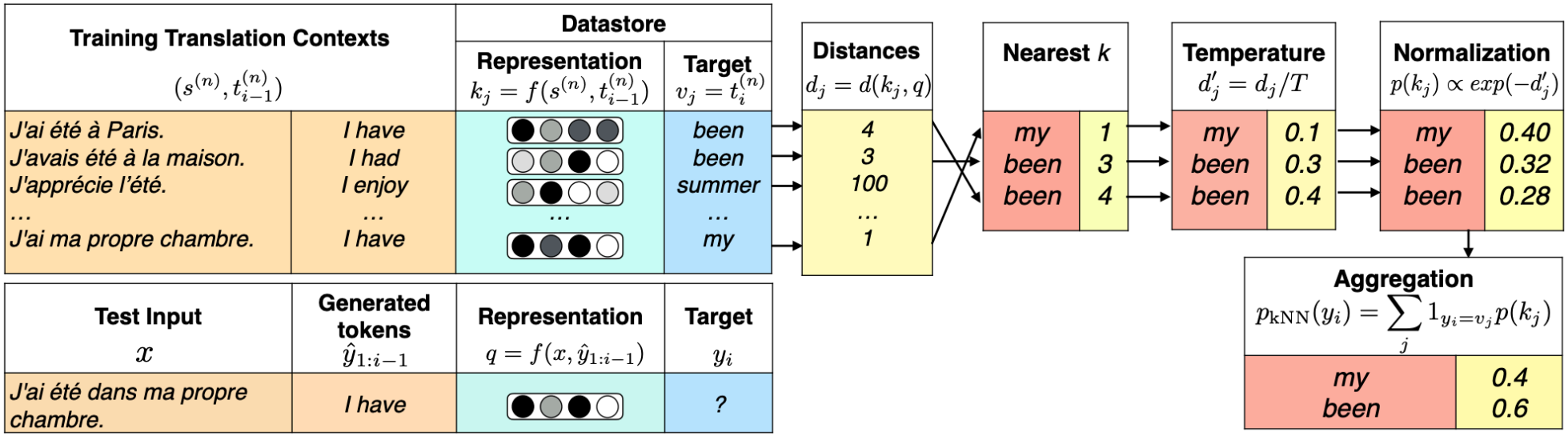


Figure from: Khandelwal et al. Nearest Neighbor Machine Translation. ICLR'2021.

Background

- Follow-up work
  - Performance enhancement
    - Jiang et al. Learning Kernel-Smoothed Machine Translation with Retrieved Examples. EMNLP-2021.
    - Zheng et al. Adaptive Nearest Neighbor Machine Translation. ACL-2021.
    - Jiang et al. Towards Robust k-Nearest-Neighbor Machine Translation. EMNLP-2022.
  - Efficiency optimization
    - Martins et al. Efficient Machine Translation Domain Adaptation. WSMNLP-2022.
    - Wang et al. Efficient Cluster-Based k-Nearest-Neighbor Machine Translation. ACL-2022.
    - Zhu et al. What Knowledge Is Needed? Towards Explainable Memory for kNN-MT Domain Adaptation. ACL-2023.

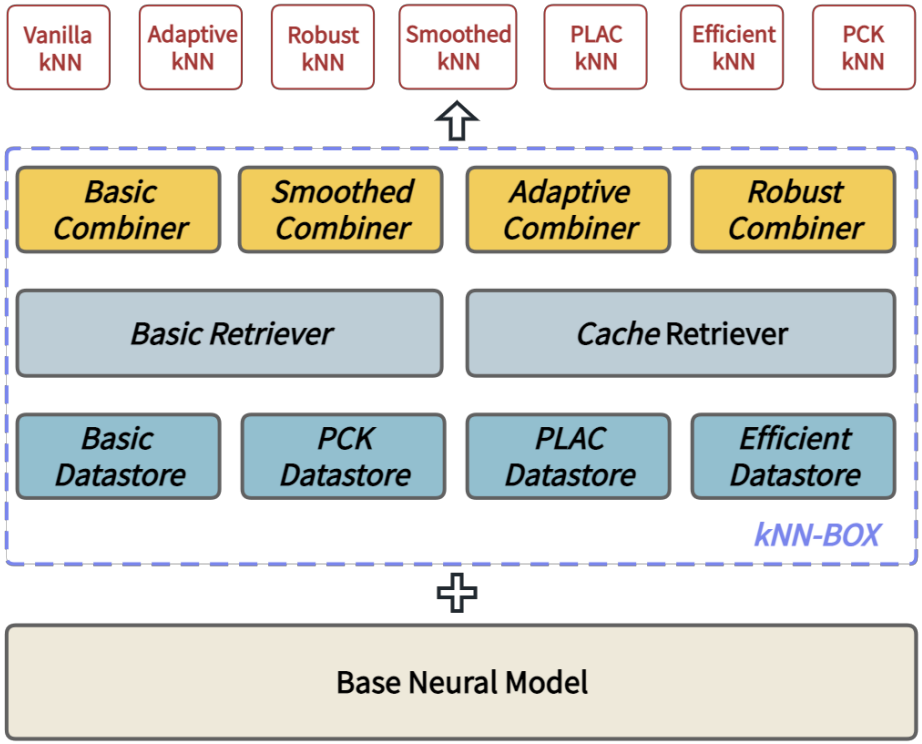
KNN-BOX Toolkit

- Motivation
  - Existing methods are implemented with diverse codebases.
  - The community can still not well understand why this paradigm works.
  - We wonder whether this approach can be applied to other seq2seq generation tasks and bring broader impact.

- Contribution
  - Developing a unified nearest neighbor generation framework.
  - Providing a GUI to show the working flow of the neural-symbolic system.
  - Demonstrating the value of this paradigm on more generation tasks.

Design of KNN-BOX

- We decompose the datastore-augmentation approach into three modules:
  - **datastore**: saving knowledge
  - **retriever**: retrieving nearest neighbors from the datastore
  - **combiner**: interpolating the output distribution of the neural model and symbolic datastore



Reproducing Existing Work

- KNN-BOX has released implementation of seven popular kNN-MT methods.
- Users can quickly reproduce existing work.

Model	Reference	Law		Medical		IT		Koran	
		Scale↓	BLEU↑	Scale↓	BLEU↑	Scale↓	BLEU↑	Scale↓	BLEU↑
Base Neural Model	Ng et al., 2019	-	45.5	100%	40.0	-	38.4	-	16.3
Vanilla *k*NN-MT	Khandelwal et al., 2021	100%	61.3	100%	54.1	100%	45.6	100%	20.4
Adaptive *k*NN-MT	Zheng et al., 2021	100%	62.9	100%	56.1	100%	47.2	100%	20.3
Smoothed *k*NN-MT	Jiang et al., 2021	100%	63.3	100%	56.8	100%	47.7	100%	19.9
Robust *k*NN-MT	Jiang et al., 2022	100%	63.6	100%	57.1	100%	48.6	100%	20.5
PCK *k*NN-MT	Wang et al., 2022	90%	62.8	90%	56.4	90%	47.4	90%	19.4
Efficient *k*NN-MT	Martins et al., 2022	57%	59.9	58%	52.3	63%	44.9	66%	19.9
PLAC *k*NN-MT	Zhu et al., 2023a	55%	62.8	55%	56.2	60%	47.0	75%	19.9

Reliable Reproduction

- We carefully compare the reproduced results with the results produced by their original implementation and find that two groups of results are well-aligned.

Model	Law	Medical	IT	Koran
Base NMT[15]	45.5	40.0	38.4	16.3
↪ *k*NN-BOX	45.5	40.0	38.4	16.3
Vanilla *k*NN-MT [16]	61.3	54.1	45.6	20.4
↪ *k*NN-BOX	61.3	54.1	45.6	20.4
Adaptive *k*NN-MT [17]	62.9	56.6	47.6	20.6
↪ *k*NN-BOX	62.9	56.1	47.2	20.3
PCK *k*NN-MT [18]	63.1	56.5	47.9	19.7
↪ *k*NN-BOX	62.8	56.4	47.4	19.4
Robust *k*NN-MT [19]	63.8	57.0	48.7	20.8
↪ *k*NN-BOX	63.6	57.1	48.6	20.5

Developing New Models

- KNN-BOX enables users to easily build a fused model, e.g., combining the most explainable datastore (PLACDATSTORE) with the strongest combiner (ROBUSTCOMBINER).

Datastore	Retriever	Combiner	Scale↓	BLEU↑
BASICDATASTORE	BASICRETRIEVER	BASICCOMBINER	100%	61.3
PCKDATASTORE	BASICRETRIEVER	ADAPTIVECOMBINER	90%	62.8
EFFICIENTDATASTORE	BASICRETRIEVER	ADAPTIVECOMBINER	57%	61.5
EFFICIENTDATASTORE	BASICRETRIEVER	ROBUSTCOMBINER	57%	61.8
PLACDATASTORE	BASICRETRIEVER	ADAPTIVECOMBINER	55%	62.8
PLACDATASTORE	BASICRETRIEVER	ROBUSTCOMBINER	55%	63.7

Visualizing Generation Process

- By running our provided script to launch a web page, users can interact with their kNN-MT system and see the visualized results.

Generation Results

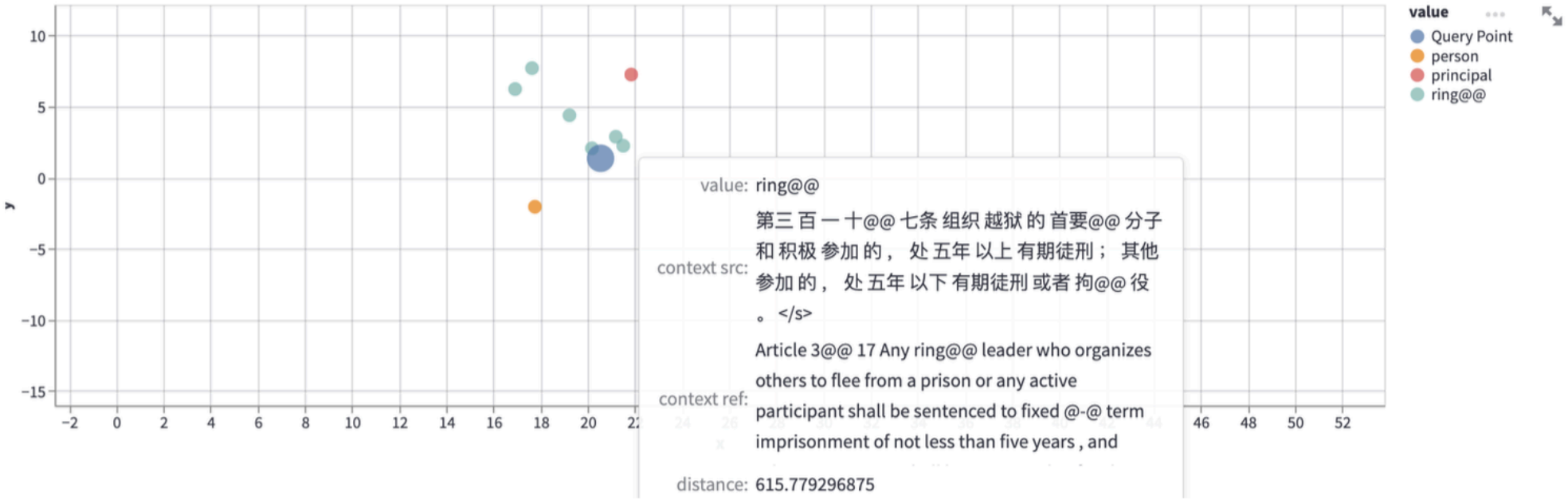
Any ringleader who organizes a jailbreak and any active participant shall be sentenced to fixed-term imprisonment of not less than five years. </s>

Any ring@@ leader who organizes a j@@ ail@@ break and any active participant shall be sentenced to fixed @-@ term imprisonment of no

	Base candidates	Base probability	kNN candidates	kNN probability
0	of	0.272	ring@@	0.775
1	ring@@	0.255	of	0.082
2	one	0.078	one	0.023
3	organization	0.019	organization	0.006
4	member	0.017	person	0.005
5	person	0.014	member	0.005
6	chief	0.012	chief	0.004
7	first	0.010	principal	0.003

Visualizing Generation Process

- When selecting on a certain nearest neighbor point, users can see the corresponding value token, translation context and query-key distance.



More Application Scenarios

- Multilingual machine translation
  - Applying kNN-BOX brings large performance improvement on all translation directions.

Directions	Model	Avg.	Cs	Da	De	Es	Fr	It	Nl	Pl	Pt	Sv
En → X	M2M-100	29.1	20.7	36.2	26.7	35.1	33.7	29.8	27.7	15.6	31.9	33.7
	+ *k*NN-BOX	**32.6**	**22.3**	**40.2**	**29.5**	**39.2**	**38.7**	**33.5**	**31.9**	**17.9**	**37.1**	**36.0**
X → En	M2M-100	33.4	27.5	40.0	31.8	36.6	35.1	33.4	31.9	21.1	38.9	37.3
	+ *k*NN-BOX	**37.7**	**31.3**	**44.5**	**37.1**	**42.0**	**40.4**	**38.4**	**36.2**	**24.9**	**41.8**	**41.0**

More Application Scenarios

- Text simplification, paraphrase generation & question generation
  - Augmenting the base neural model with kNN-BOX brings performance enhancement in all three tasks.

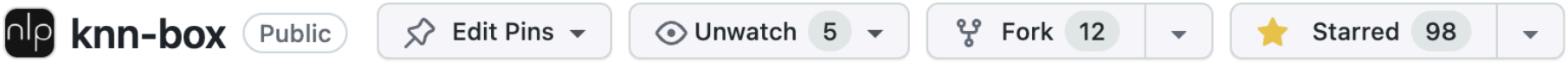
Task	Dataset	Metric	Base Model	*k*NN-BOX
Text Simplification	Wiki-Auto	SARI	38.6	**39.4**
	Newsela-Auto	SARI	35.8	**38.2**
Paraphrase Generation	QQP	BLEU	28.4	**29.5**
Question Generation	Quasar-T	BLEU	9.6	**15.7**

Quick Impact

- kNN-BOX has been used as the backbone of several works.
  - Liu et al. kNN-TL: k-Nearest-Neighbor Transfer Learning for Low-Resource Neural Machine Translation. ACL-2023.
  - Li et al. Revisiting Source Context in Nearest Neighbor Machine Translation. EMNLP-2023.
  - Zhang et al. Syntax-Aware Retrieval Augmented Code Generation. EMNLP-2023.
  - Zhang et al. NNOSE: Nearest Neighbor Occupational Skill Extraction. EACL-2024.

Conclusion

- We develop an open-sourced toolkit kNN-BOX for nearest neighbor generation.
  - quickly reproducing existing works
  - flexibly fusing advanced techniques
  - visually analyzing generation process



https://github.com/NJUNLP/knn-box