What Knowledge Is Needed? Towards Explainable Memory for kNN-MT Domain Adaptation

Wenhao Zhu[1,2], Shujian Huang[1,2], Yunzhe Lv[1,2], Xin Zheng[1,2], Jiajun Chen[1,2]

National Key Laboratory for Novel Software Technology, Nanjing University

Collaborative Innovation Center of Novel Software Technology and Industrialization

Motivation

- kNN-MT incorporates the symbolic datstore to assist the neural model, which usually saves all target language token occurences in the parallel corpus.
- The constructed datastore is usually large and possibly redundant.

translation context $(X, Y_{<t})$		hidden state $h(X, Y_{<t})$	target token y_t
Wie wirkt Penizillin?	<bos>		How
Wie wirkt Penizillin?	<bos> How		Does
Wie wirkt Penizillin?	<bos> How does		Penicillin
Wie wirkt Penizillin?	<bos> How does Penicillin		work
Wie wirkt Penizillin?	<bos> How does Penicillin work		?
Wie wirkt Penizillin?	<bos> How does Penicillin work ?		<eos>

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

key-value datastore

What Knowledge Does the Neural Model Need?

- The relationship between NMT model and symbolic datastore is unclear.

- Intuitively, the pre-trained NMT model only needs knowledge that remedies its weakness.

- We propose to explore this issue from the point of “local correctness”
  - translation correctness for a single entry (entry correctness)
  - Translation correctness for a given neighborhood (neighborhood correctness).

Local Correctness

- Entry Correctness
  - Entry correctness describes whether the NMT model could make correct translation for a specific entry.
  - It can be evaluated by comparing target token and prediction token:

translation context $(X, Y_{<t})$		hidden state $h(X, Y_{<t})$	target token y_t	predict token $\hat{y}_t$	
Wie wirkt Penizillin ?	<bos>		How	How	known
Wie wirkt Penizillin ?	<bos> How		Does	Does	known
Wie wirkt Penizillin ?	<bos> How does		Penizillin	Cyanokit	unknown
Wie wirkt Penizillin ?	<bos> How does Penicillin		work	works	unknown
Wie wirkt Penizillin ?	<bos> How does Penicillin work		?	?	known
Wie wirkt Penizillin ?	<bos> How does Penicillin work ?		<eos>	<eos>	known

check

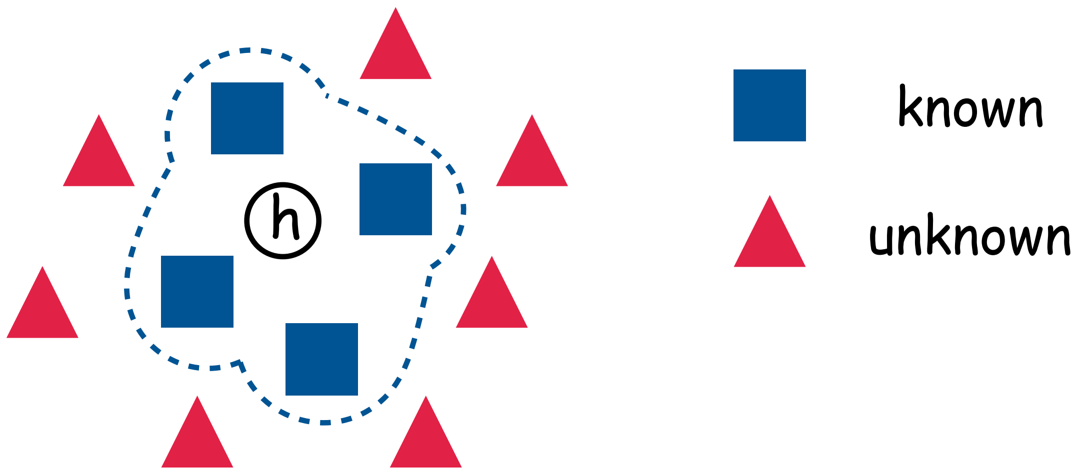
Local Correctness

- Neighborhood Correctness
  - Neighborhood correctness evaluates the NMT model's prediction on a neighborhood in the representation space.
  - Knowledge margin is proposed as the metric.

knowledge margin

$$km(h) = \arg\max_t \forall (h^j, y^j) \in \mathcal{N}_t(h) \text{ is known}$$

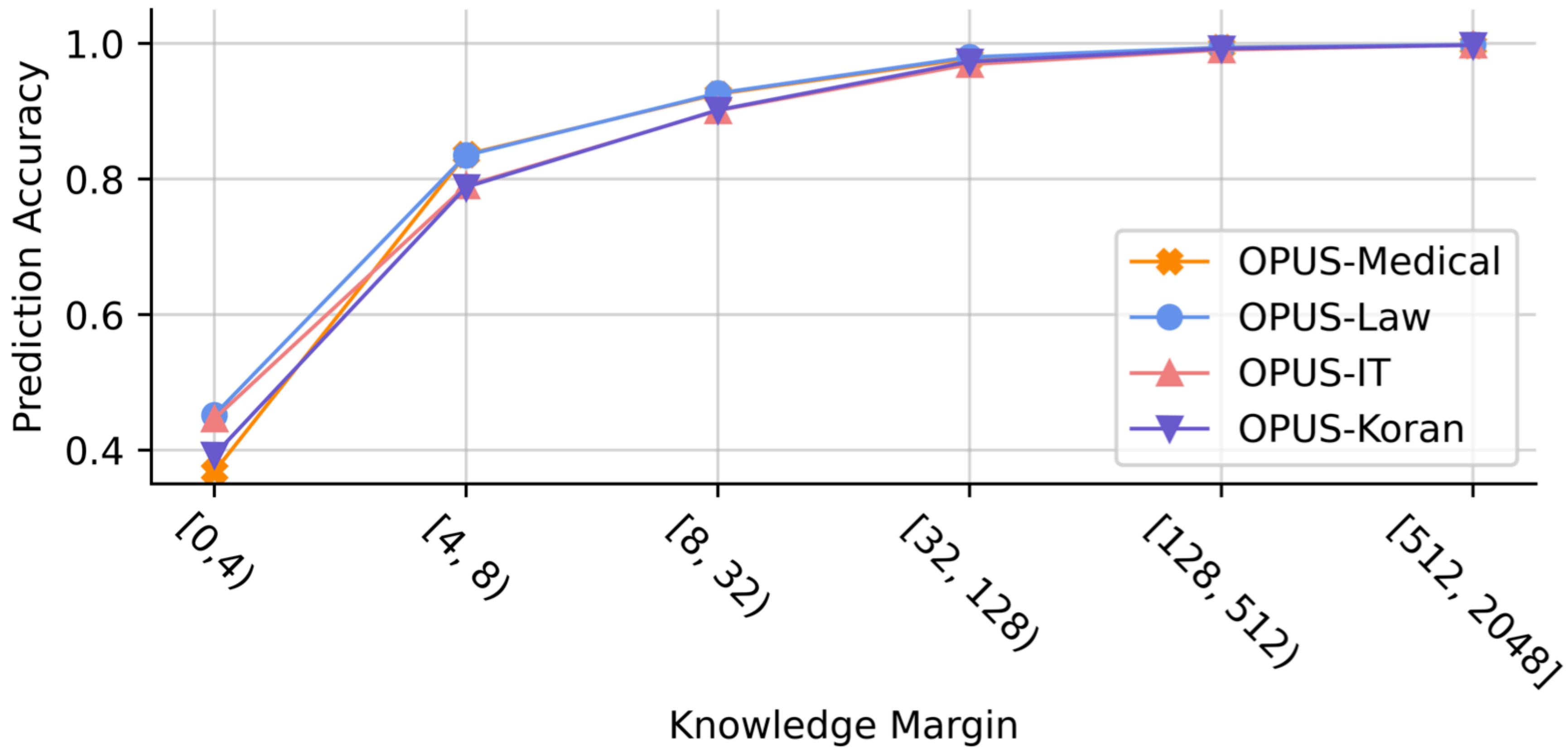
$$km(h) = 4$$



Intuitively, km is the maximum size of the neighborhood of the entry h where the NMT could make correct translation

Local Correctness

- Knownledge margin value can reflect the capability of the NMT model.



Local Correctness

- Understand the role of different datastore entries.
  - Entries with small km: NMT model tends to fail when context are similar but different. helpful
  - Entries with large km: NMT model generalizes well on these entries. less helpful
- PLAC: Pruning with LocAl Correcness

Algorithm 1 Datastore Pruning by PLAC

Input: datastore $\mathcal{D}$, the *knowledge margin* threshold k_p, the pruning ratio r
Output: pruned datastore $\mathcal{D}$

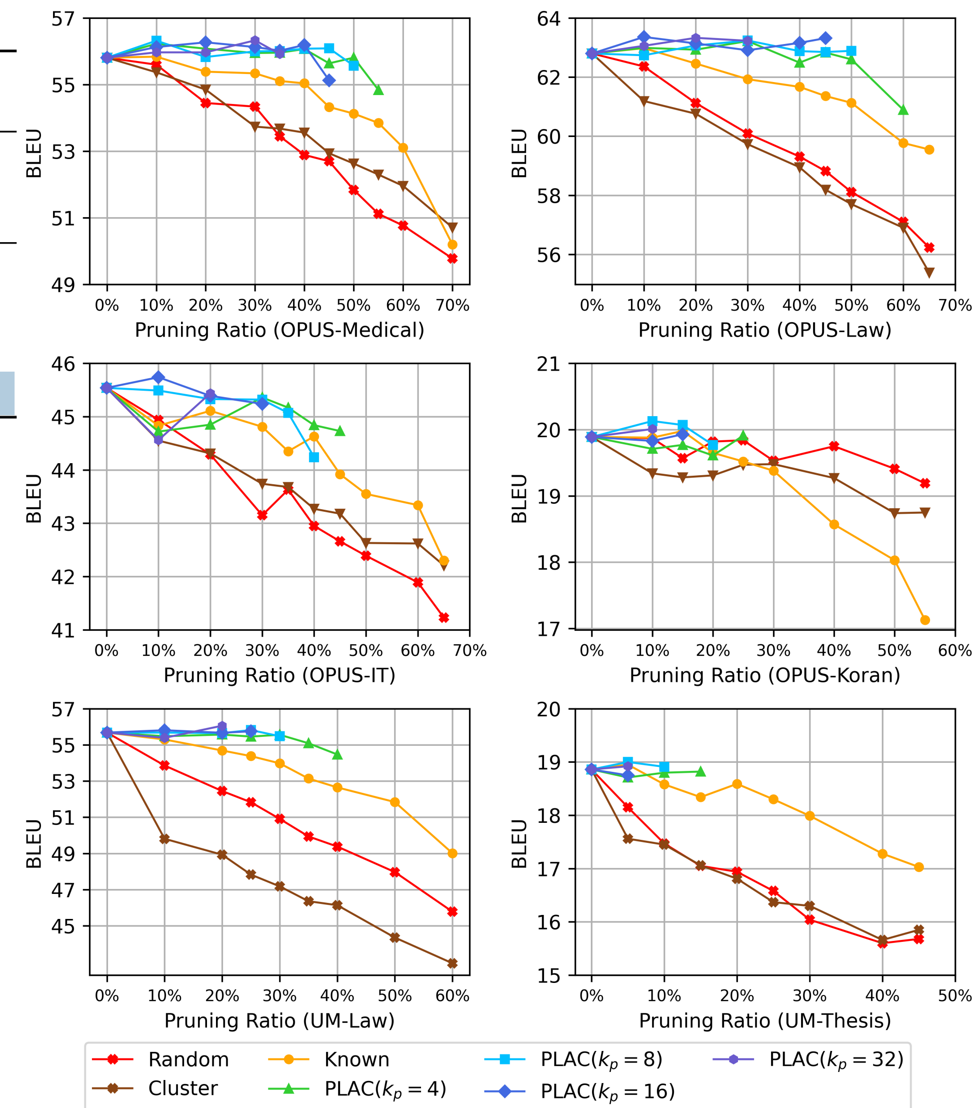
$candidates \leftarrow \emptyset$ ▷ step 1: collect
for each entry (h, y) in $\mathcal{D}$ **do**
  if $km(h) \geq k_p$ **then**:
    $candidates \leftarrow candidates \cup (h, y)$
  end if
end for
repeat ▷ step 2: drop
  randomly select entry (h, y) from $candidates$
  remove (h, y) from $\mathcal{D}$
until pruning ratio r is satisfied
return $\mathcal{D}$

Experiment Results

- Pruning with local correctness (PLAC) cuts off 25%-45% datastore entries while achieve comparable performance
  - Previous pruning method (40% -1.4 BLEU, 10% -0.9 BLEU)

	OPUS-Medical			OPUS-Law			OPUS-IT			OPUS-Koran		
	Ratio	BLEU↑	COMET↑	Ratio	BLEU↑	COMET↑	Ratio	BLEU↑	COMET↑	Ratio	BLEU↑	COMET↑
Base	-	39.73	0.4665	-	45.68	0.5761	-	37.94	0.3862	-	16.37	-0.0097
Finetune	-	58.09	0.5725	-	62.67	0.6849	-	49.08	0.6343	-	22.40	0.0551
Adaptive *k*NN	0%	57.98	0.5801	0%	63.53	0.7033	0%	48.39	0.5694	0%	20.67	0.0364
Random	45%	54.08*	0.5677*	45%	58.69*	0.6690*	40%	45.54*	0.5314*	25%	20.36	0.0434
Cluster	45%	53.31*	0.5689*	45%	58.68*	0.6779*	40%	45.80*	0.5788	25%	20.04*	0.0410*
Known	45%	56.44*	0.5691*	45%	61.61*	0.6885*	40%	45.93*	0.5563*	25%	20.35	0.0338
All Known	73%	42.73*	0.4926*	66%	51.90*	0.6200*	69%	40.93*	0.4604*	56%	17.76*	0.0008*
PLAC (ours)	45%	57.66	0.5773	45%	63.22	0.6953*	40%	48.22	0.5560	25%	20.96	0.0442

	UM-Law			UM-Thesis		
	Ratio	BLEU↑	COMET↑	Ratio	BLEU↑	COMET↑
Base	-	30.36	0.3857	-	13.13	-0.0442
Finetune	-	58.55	0.6019	-	17.46	-0.0262
Adaptive *k*NN	0%	58.64	0.6017	0%	17.49	-0.0146
Random	30%	53.78*	0.5661*	15%	16.14*	-0.0280*
Cluster	30%	49.65*	0.5274*	15%	15.73*	-0.0419*
Known	30%	56.92*	0.5762*	15%	17.25	-0.0143
All Known	63%	46.45*	0.4720*	47%	15.33*	-0.0525*
PLAC (ours)	30%	58.65	0.6056	15%	17.52	-0.0122



NMT: winner model of WMT19 De-En news translation task
Dataset: OPUS, UM

Conclusion

- We analyze the local correctness of the neural model's predictions to identify the conditions where the neural model may fail.

- We find that the NMT model often fails when the knowledge margin is small.

- We can safely prune the datastore with the proposed PLAC method, validating our findings about local correctness and translation failures.