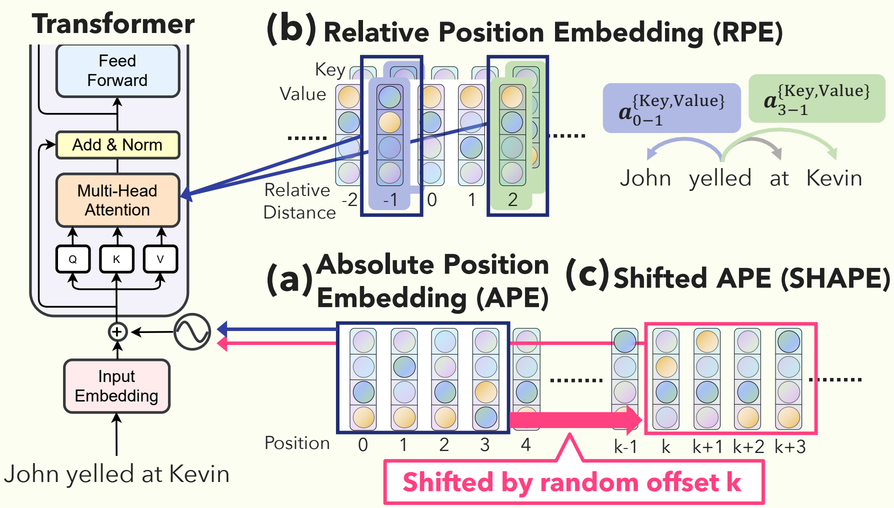
# SHAPE: Shifted Absolute Position Embedding for Transformers


Shun Kiyono[1,2], Sosuke Kobayashi[2,3], Jun Suzuki[2,1], Kentaro Inui[2,1]   [1]RIKEN   [2]Tohoku University   [3]Preferred Networks, Inc.


## Overview: building absolute position embedding that can extrapolate and shift-invariant




**i.e. generalize to lengths unseen during training**

### Summary of Position Representations

| | APE | RPE | SHAPE |
|---|---|---|---|
| Performance | 😰 | 😀 | 😀 |
| Speed | 😀 | 😰 | 😀 |
| Implementation | 😀 | 😰 | 😀 |

**SHAPE Implementation: a single line**

```
pos_embed(pos_idx)
```

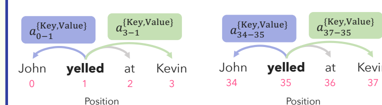
```
pos_idx += self.training * torch.randint(0, K)
pos_embed(pos_idx)
```

**Idea: Bring RPE's success to APE**

Shift invariance: spatial shift does not change function's output

RPE is shift-invariant

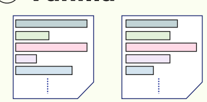
→ $a_{0-1}^{\{Key,Value\}}$ is same as $a_{34-35}^{\{Key,Value\}}$




Random shift in APE can prevent the use of absolute positions
→ Model learns shift invariance?

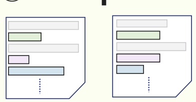
## Experiment on Machine Translation (WMT EnDe)

### ① Vanilla



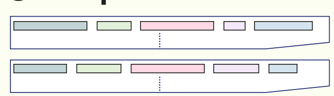
- Standard setting for MT
- Sanity check of baseline performance

### ② Extrapolate



- Remove sequence longer than 50 tokens
- Evaluate performance on unseen length

### ③ Interpolate



- Concatenate neighboring sequences
- Tokens are more infrequent at given position

## Experimental Result and Observations

| Dataset | Model | Valid | Test | Speed |
|---|---|---|---|---|
| VANILLA | APE† | 23.61 | 30.46 | x1.00 |
| | RPE† | 23.67 | 30.54 | x0.91 |
| | SHAPE† | 23.63 | 30.49 | x1.01 |
| EXTRAPOLATE | APE | 22.18 | 29.22 | x1.00 |
| | RPE | 22.97 | 29.86 | x0.91 |
| | SHAPE | 22.96 | 29.80 | x0.99 |
| INTERPOLATE | APE | 31.40 | 38.23 | x1.00 |
| | RPE* | - | - | - |
| | SHAPE | 32.50 | 39.09 | x0.99 |

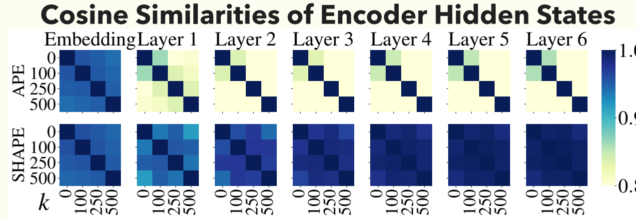
- All models are comparable
- **SHAPE has no risk of performance drop**

- **Both RPE and SHAPE outperform APE**
- **SHAPE is as fast as APE** while RPE is not

- RPE is prohibitively slow
- **SHAPE outperforms APE**

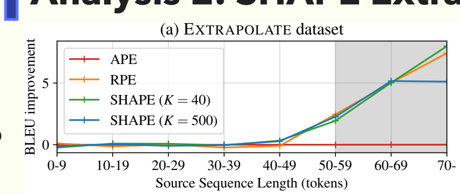
## Analysis 1: SHAPE is Shift-invariant

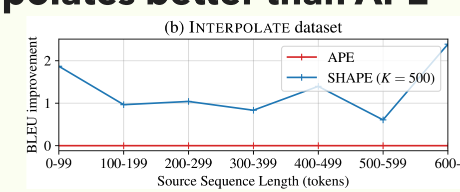
### Cosine Similarities of Encoder Hidden States




SHAPE has high similarities regardless of offset $k$

## Analysis 2: SHAPE Extrapolates better than APE




- SHAPE outperforms APE in gray no-training-data region → **better extrapolation than APE**

- **SHAPE outperforms APE**
- Data augmentation effect?