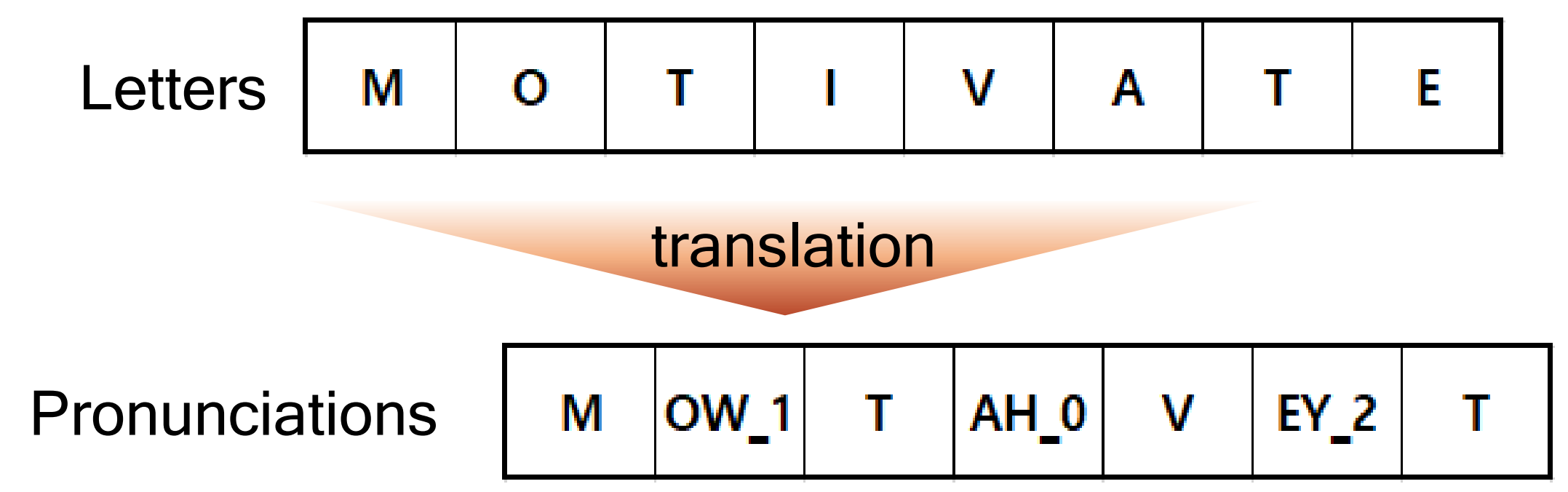
CONVOLUTIONAL SEQUENCE TO SEQUENCE MODEL WITH NON-SEQUENTIAL GREEDY DECODING FOR GRAPHEME TO PHONEME CONVERSION

Moon-jung Chae* jjamjung@snu.ac.kr
Kyubyong Park† kbpark.linguist@gmail.com
Jinhyun Bang* jinhyun95@snu.ac.kr
Soobin Suh* soobin3230@snu.ac.kr
Jonghyuk Park* chico2121@snu.ac.kr
Namju Kim† buriburisuri@gmail.com
Jonghun Park* jonghun@snu.ac.kr

*Department of Industrial Engineering & Center for Superintelligence, Seoul National University

†KakaoBrain corporation

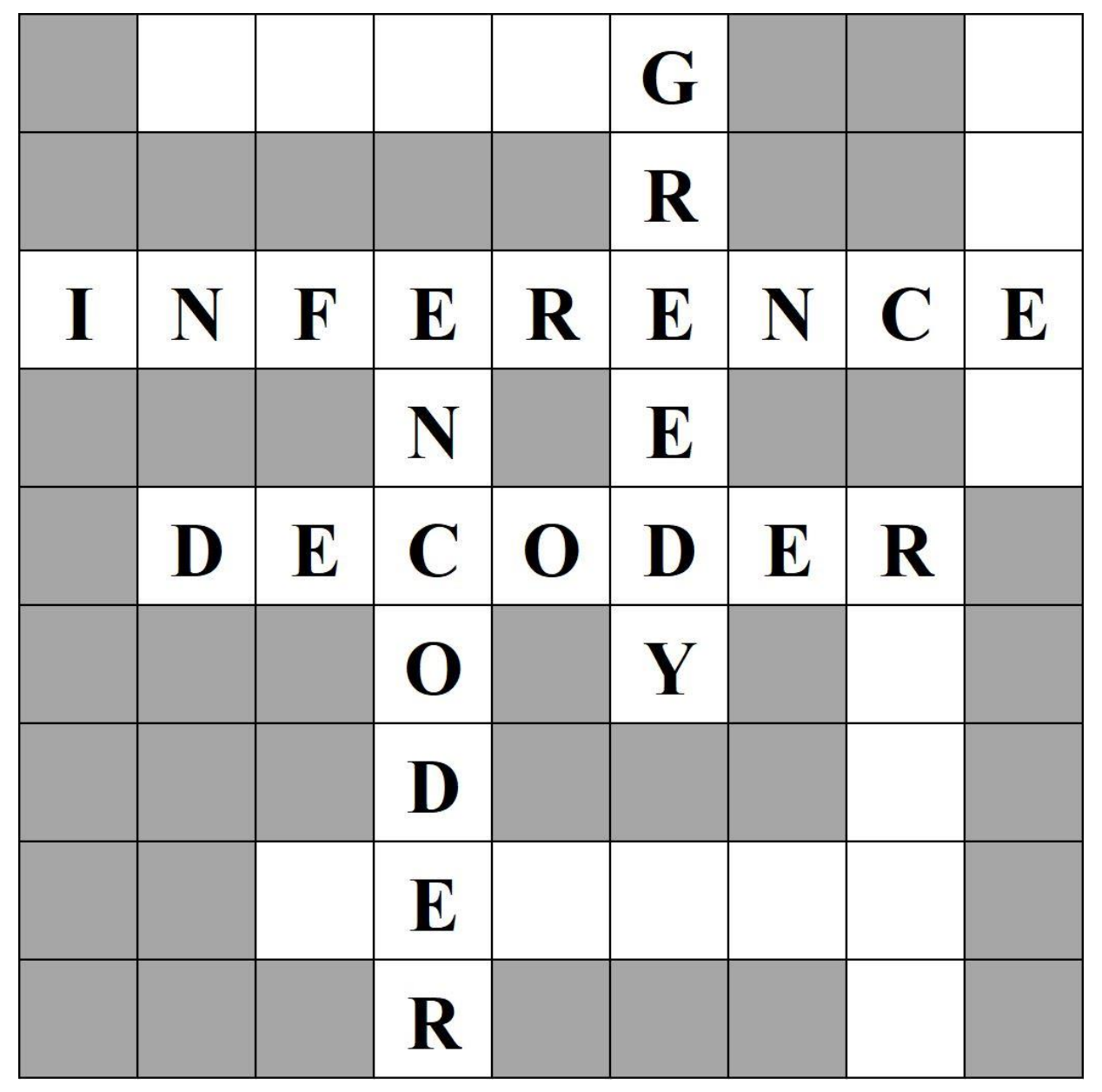
Grapheme-to-phoneme (G2P) Task



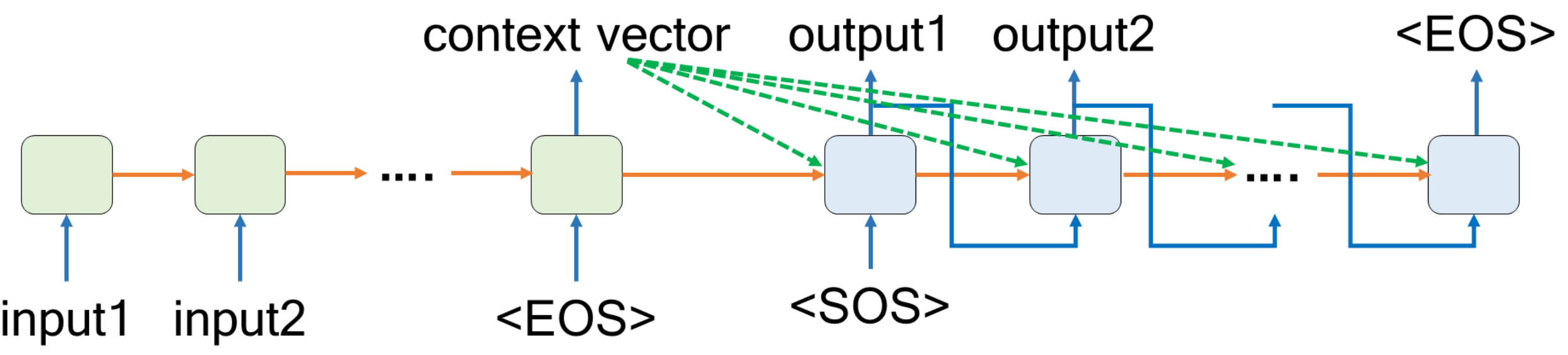
- G2P models have been frequently employed in text-to-speech (TTS) and automatic speech recognition (ASR) systems.
- Convolutional sequence-to-sequence models have not been applied to G2P problem yet.

Non-sequential Greedy Decoding (NSGD)

When you play a crossword puzzle, which blank will you fill in first?



- **It is a good strategy to fill in the easiest blank first**, referring to the hints given so far.
- According to this greedy strategy, the easy parts filled in earlier can be used as hints later.

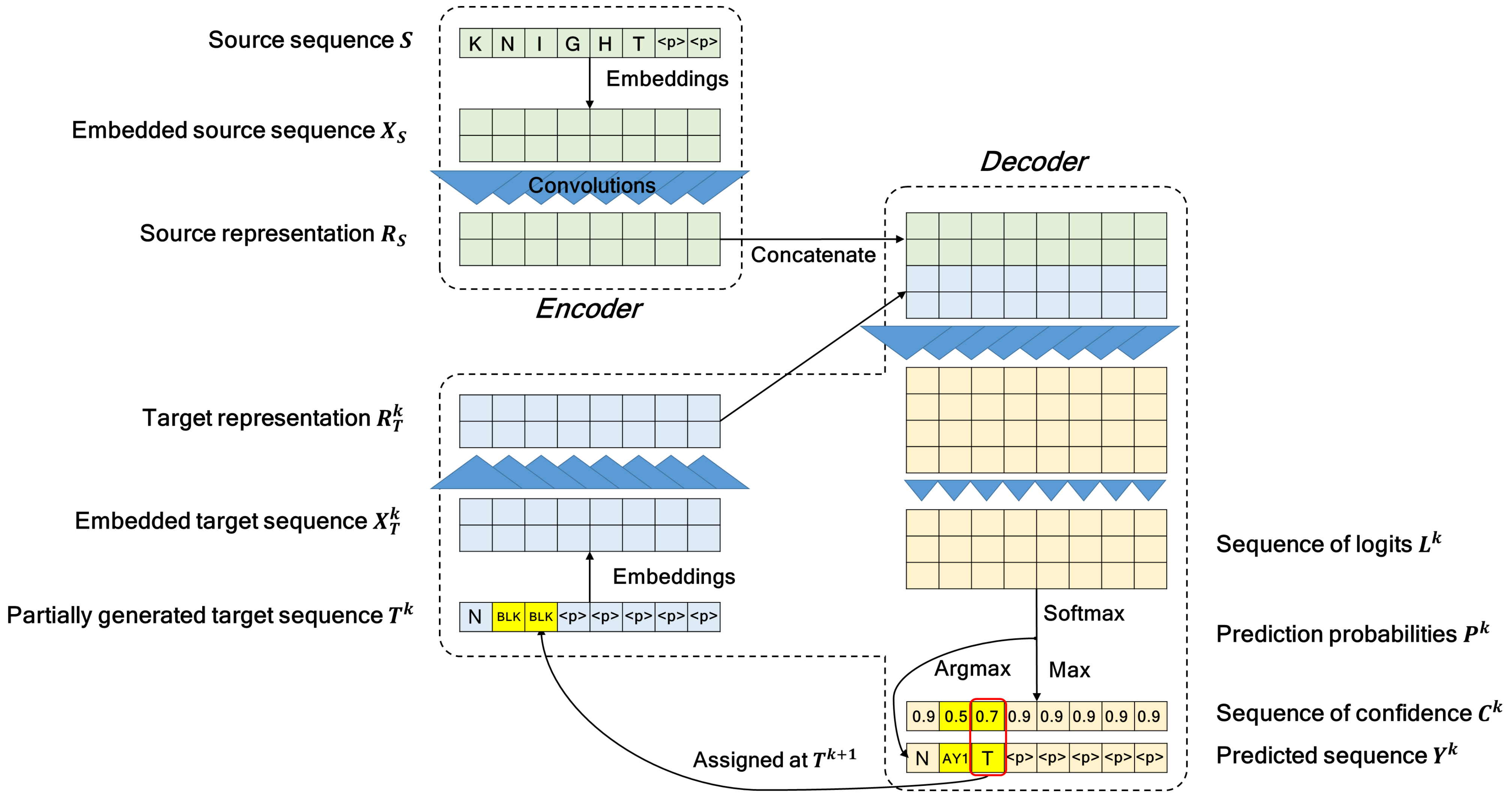


- **Most encoder-decoder models proposed in the past carry out decoding sequentially**, one step at a time from left to right, and use the outputs from the previous steps as decoder inputs.
- In some cases, the beginning of the sequence may be the most difficult part to infer, and **inaccurate inference in the beginning can negatively affect what follows, which may result in serious compounding errors.**

Key idea: decoding the easy part first

- **We propose a non-sequential greedy decoding (NSGD) method which is a generalized version of the greedy decoding.**
- Our method considers not only which token to select next but also which position to consider at each inference step.

Convolutional sequence-to-sequence model with NSGD

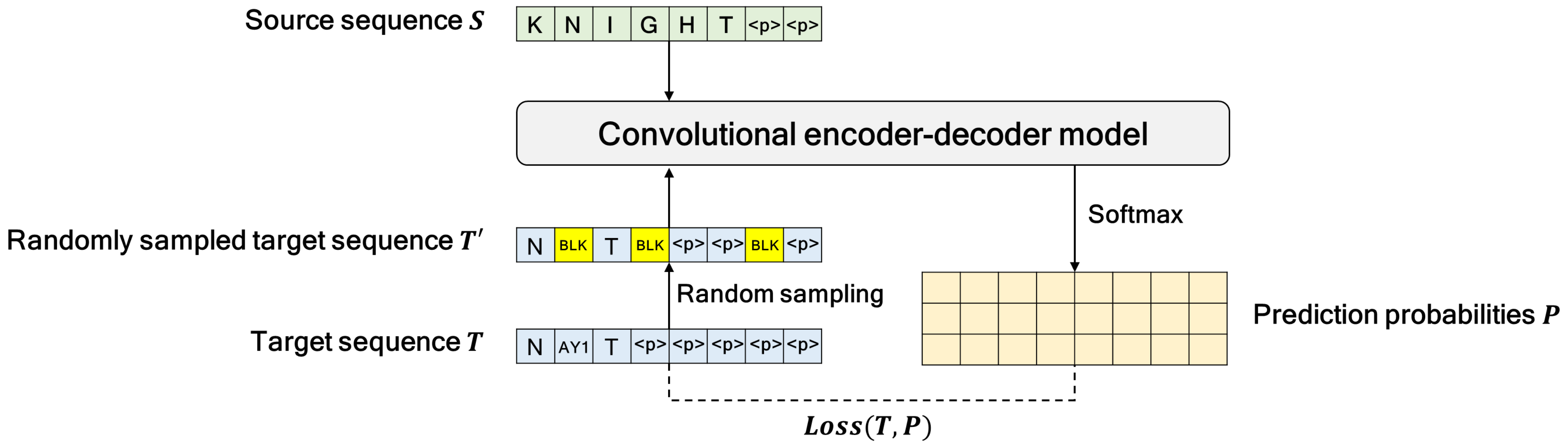

Source sequence S
K N I G H T <p> <p>
Embeddings
Embedded source sequence X_S
Convolutions
Source representation R_S
Concatenate
Encoder
Decoder
Target representation R_T^k
Embedded target sequence X_T^k
Embeddings
Partially generated target sequence T^k
N BLK BLK <p> <p> <p> <p> <p>
Sequence of logits L^k
Softmax
Prediction probabilities P^k
Argmax
Max
Sequence of confidence C^k
0.9 0.5 0.7 0.9 0.9 0.9 0.9 0.9
Predicted sequence Y^k
N AY1 T <p> <p> <p> <p> <p>
Assigned at T^{k+1}

How the proposed model works

- Our convolutional encoder-decoder model uses two inputs, source sequence $S=(s_1,\dots,s_N)$ and target sequence $T=(t_1,\dots,t_N)$ of N tokens each.
- The model outputs prediction probabilities $P=(P_1,\dots,P_N)$.
- **With NSGD, the model iteratively infers the most likely part among the candidate positions of T that are not inferred yet.**
- NSGD initially starts with T^0, and updates T^k to T^{k+1} until the fully generated target sequence T^N is obtained.

Algorithm 1 Inference procedure with NSGD

```
Require: sequence-to-sequence model $G_\theta$ with weights $\theta$; source
    sequence $S$.
 1: Initialize target sequence $T^0$ with blank tokens and set $I=\emptyset$.
 2: for $k \in \{0,\dots,N-1\}$ do
 3:     $T^{k+1} \leftarrow T^k$.
 4:     Generate $P^k$ using $G_\theta$ with inputs $S$ and $T^k$.
 5:     Compute $Y^k=(y_1^k,\dots,y_N^k)$ and $C^k=(c_1^k,\dots,c_N^k)$ from
        $P^k$ using argmax and max, respectively.
 6:     for $n \in \{1,\dots,N\}$ do
 7:         if $n \in I$ then
 8:             $c_n^k \leftarrow 0$
 9:         end if
10:     end for
11:     $n^* \leftarrow \operatorname{argmax}_n (C^k)$.
12:     $t_{n^*}^{k+1} \leftarrow y_{n^*}^k$.
13:     $I \leftarrow I \cup \{n^*\}$.
14: end for
15: return $T^N$.
```


Source sequence S
K N I G H T <p> <p>
Convolutional encoder-decoder model
Softmax
Randomly sampled target sequence T'
N BLK T BLK <p> <p> BLK <p>
Random sampling
Target sequence T
N AY1 T <p> <p> <p> <p> <p>
Prediction probabilities P
$Loss(T,P)$

Training procedure

- We use random sampling function $r: T \rightarrow T'$ which replaces a random subset of ground truth tokens in T with blank tokens.
- Using r, **we generate randomly sampled target sequence T' and feed it to the model.**

Algorithm 2 Training procedure with NSGD

```
Require: sequence-to-sequence model $G_\theta$ with weights $\theta$; source
    sequence $S$; target sequence $T$
1: Initialize $G_\theta$ with random weights $\theta$.
2: while not converged do
3:     Generate $T'$ from $T$ with $r$ following (1).
4:     Generate $P$ using $G_\theta$ with inputs $S$ and $T'$.
5:     Update $\theta$ to minimize (2)
6: end while
```

Experiments

Dataset

- We mad three versions of datasets from CMUDict US English dataset.

Dataset	Multiple pronunciations	Stress markings	Number of instances
CMUDict-MS	Kept	Kept	133,853
CMUDict-M	Kept	Removed	133,853
CMUDict-S	Removed	Kept	116,919

Results

Model	CMUDict-MS		CMUDict-M		CMUDict-S	
	PER (%)	WER (%)	PER (%)	WER (%)	PER (%)	WER (%)
Encoder-decoder + attention	8.00 ± 0.11	30.42 ± 0.48	5.91 ± 0.07	25.19 ± 0.20	8.00 ± 0.11	30.42 ± 0.48
Encoder-decoder + attention*	7.63 ± 0.08	28.61 ± 0.37	5.72 ± 0.10	24.77 ± 0.38	7.88 ± 0.16	28.89 ± 0.41
Proposed model	$\mathbf{7.25 \pm 0.07}$	$\mathbf{28.42 \pm 0.22}$	$\mathbf{5.58 \pm 0.04}$	$\mathbf{24.10 \pm 0.19}$	$\mathbf{7.44 \pm 0.06}$	$\mathbf{28.87 \pm 0.26}$

- ± indicates the standard deviation across 5 training runs of the model.
- The baseline model marked as * was reported as the state-of-the-art.
- **The proposed model shows the best performances in terms of both PER (phoneme error rate) and WER (word error rate)** compared to the baselines including the state-of-the-art model.
- **The proposed model shows more stable results with lower standard deviations of error rates** than the baselines.

Decoding example

Source of sequence		Acquired phoneme sequence									
Our model with NSGD	$k=1$	○	○	○	○	○	○	○	○	○	●
	$k=2$	○	○	○	○	○	○	○	○	●	●
	$k=3$	○	○	t	○	○	○	○	○	●	●
	$k=4$	○	n	t	○	○	○	○	○	●	●
	$k=5$	○	n	t	r	○	○	○	○	●	●
	$k=6$	○	n	t	r	○	○	○	●	●	●
	$k=7$	○	n	t	r	○	p	○	●	●	●
	$k=8$	○	n	t	r	○	p	●	●	●	●
	$k=9$	ih0	n	t	r	○	p	●	●	●	●
	$k=10$	ih0	n	t	r	ae1	p	●	●	●	●
with sequential greedy decoding		eh1	n	t	r	ah0	p	●	●	●	●
Ground truth		ih0	n	t	r	ae1	p	●	●	●	●

- ◦ and • are symbols representing blank and padding tokens respectively.
- Among the phonemes, some vowels include digits representing stress.
- Grey cells mean incorrect prediction results.
- **The proposed model postpones the inference on vowel phonemes which is the most difficult part, and uses all the rest of the phonemes inferred so far as its input.**

Our Main Result

- We proposed a **non-sequential greedy decoding method (NSGD) that generalizes traditional greedy decoding** and two algorithms for inference and training.
- We also proposed a **fully convolutional encoder-decoder model for NSGD**.
- We were able to show the effectiveness of the proposed model and the decoding method by **achieving the state-of-the-art performances on the G2P task**.