OpenNMT: An Introduction

Alexander Rush @srush, Harvard University

OpenNMT Workshop Paris, February 2018

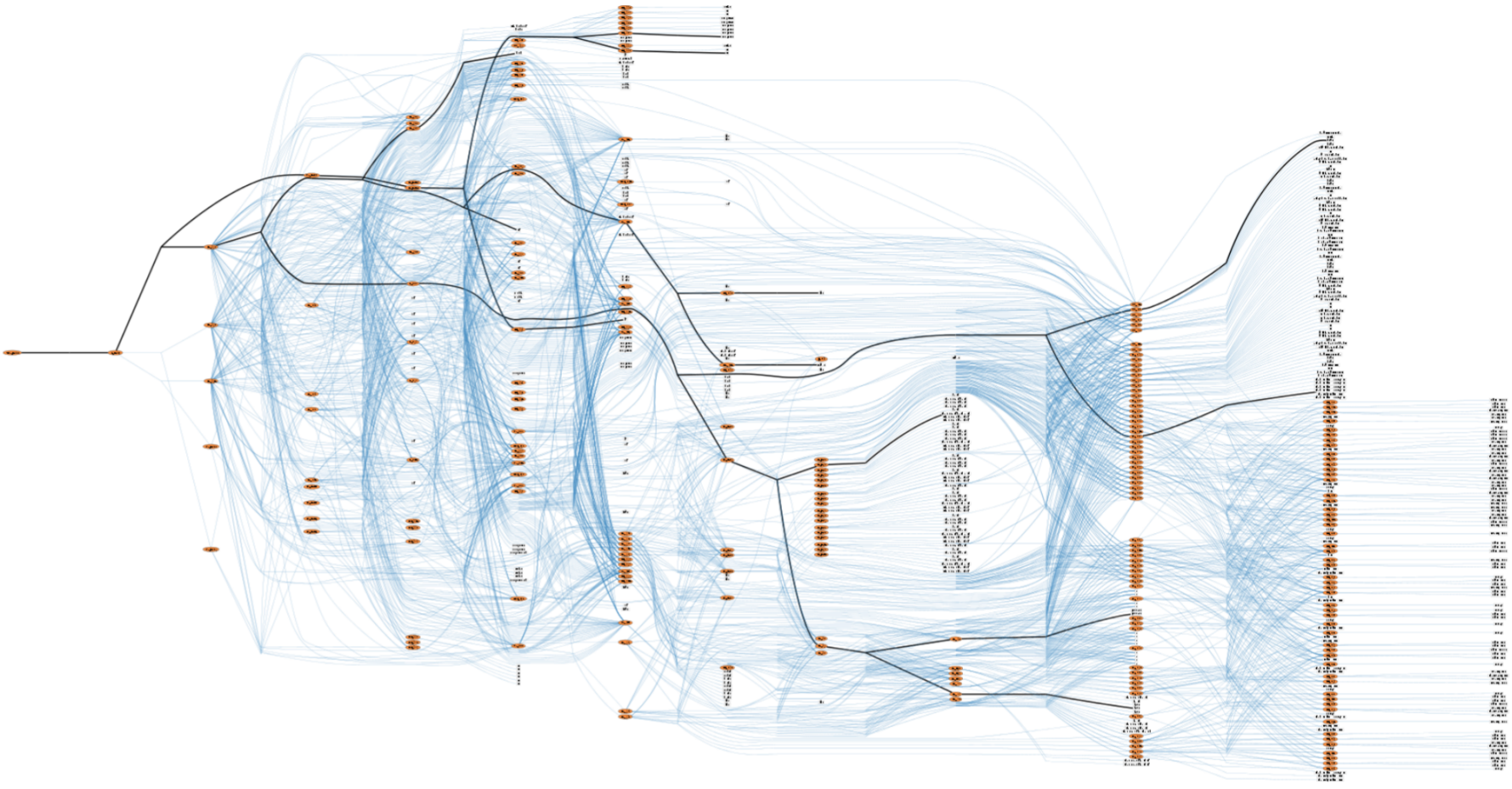
HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

A Short (Personal) History of OpenNMT

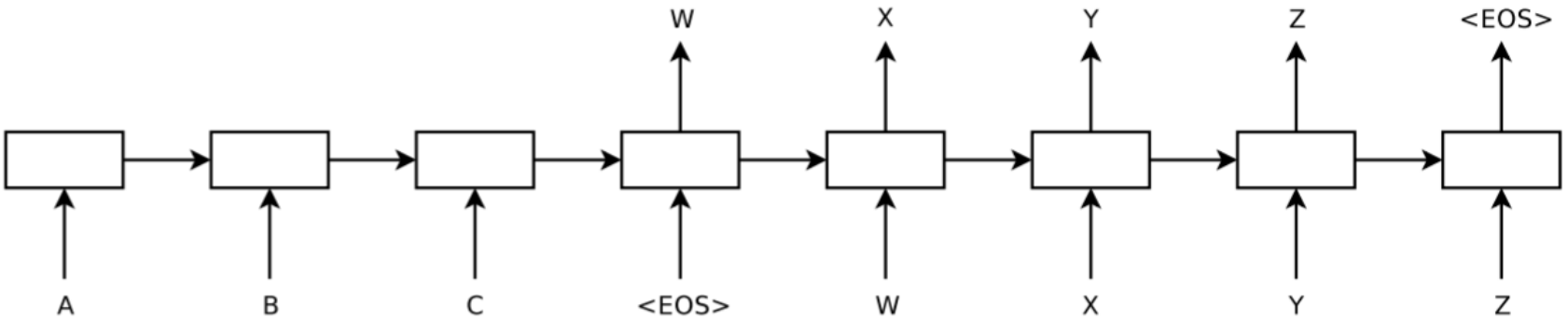
2013: Syntactic Translation

载入 黛 妃 死 因 调查 资料 的 两 台 手 提 电脑 遭 窃

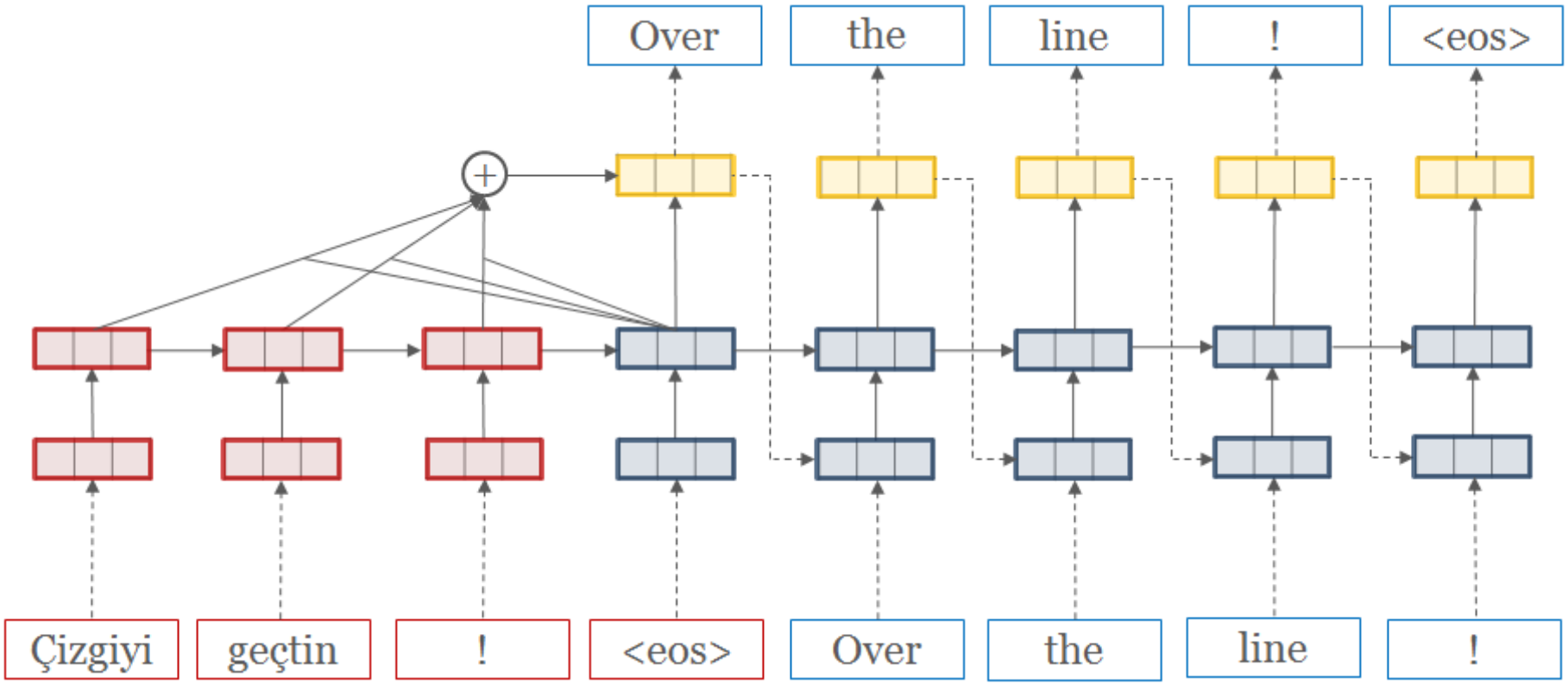
Two laptops with information on the cause of Princess Diana's death were stolen



2014: Advent of Neural MT



2015: Developing Seq2seq-attn



2015: Developing Seq2seq-attn

Our Group (circa 2015):

2015: Developing Seq2seq-attn

github.com/harvardnlp/seq2seq-attn

README.md

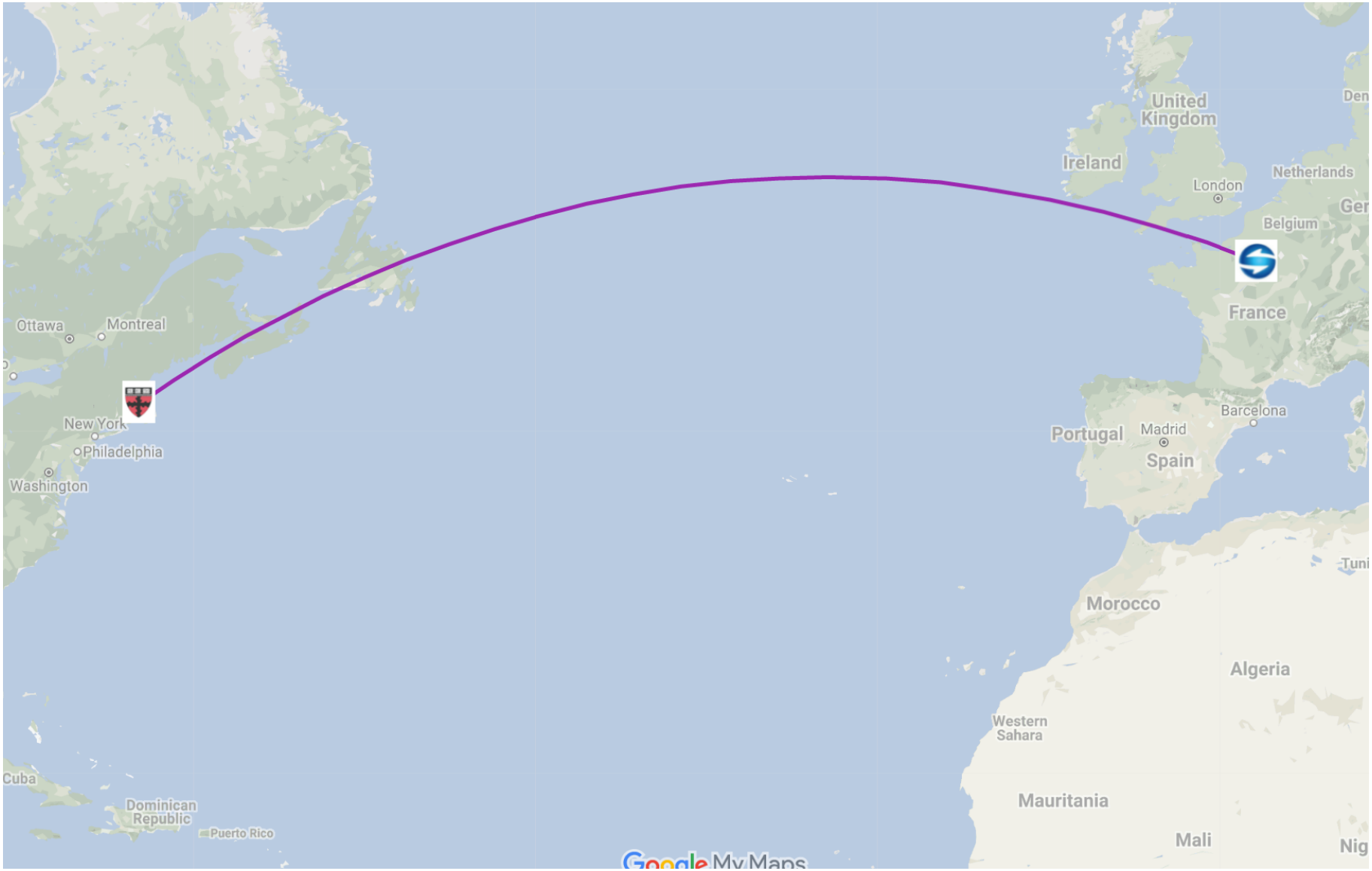
Sequence-to-Sequence Learning with Attentional Neural Networks

Torch implementation of a standard sequence-to-sequence model with attention where the encoder-decoder are LSTMs. Also has the option to use characters (instead of input word embeddings) by running a convolutional neural network followed by a highway network over character embeddings to use as inputs.

The attention model is from Effective Approaches to Attention-based Neural Machine Translation, Luong et al. EMNLP 2015. We use the *global-attention* model with the *input-feeding* approach from the paper.

The character model is from Character-Aware Neural Language Models, Kim et al. AAAI 2016.

2016: From Research to Production



2016: From Research to Production

SYSTRAN's Pure Neural Machine Translation Systems

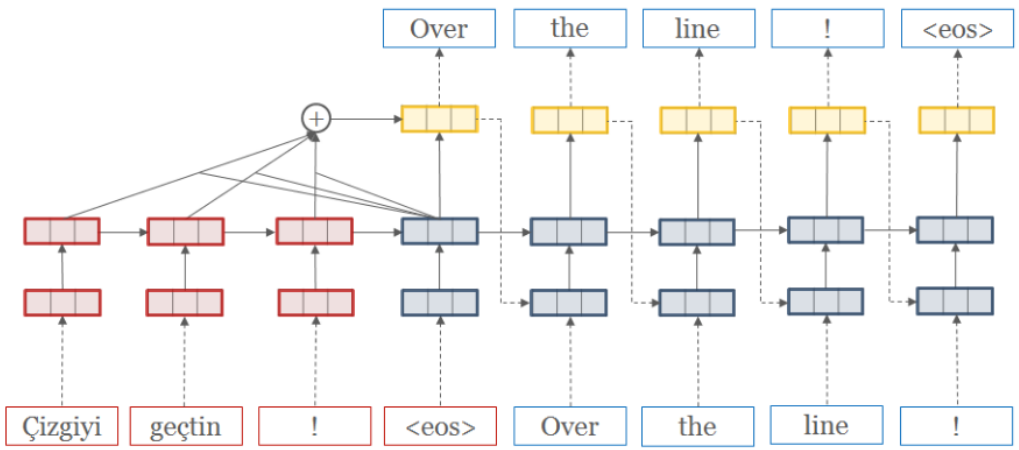
Josep Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart
Egor Akhanov, Patrice Brunelle, Aurélien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss
Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux
Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal
Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou, Peter Zoldan

SYSTRAN
firstname.lastname@systrangroup.com

Abstract **1 Introduction**

Home

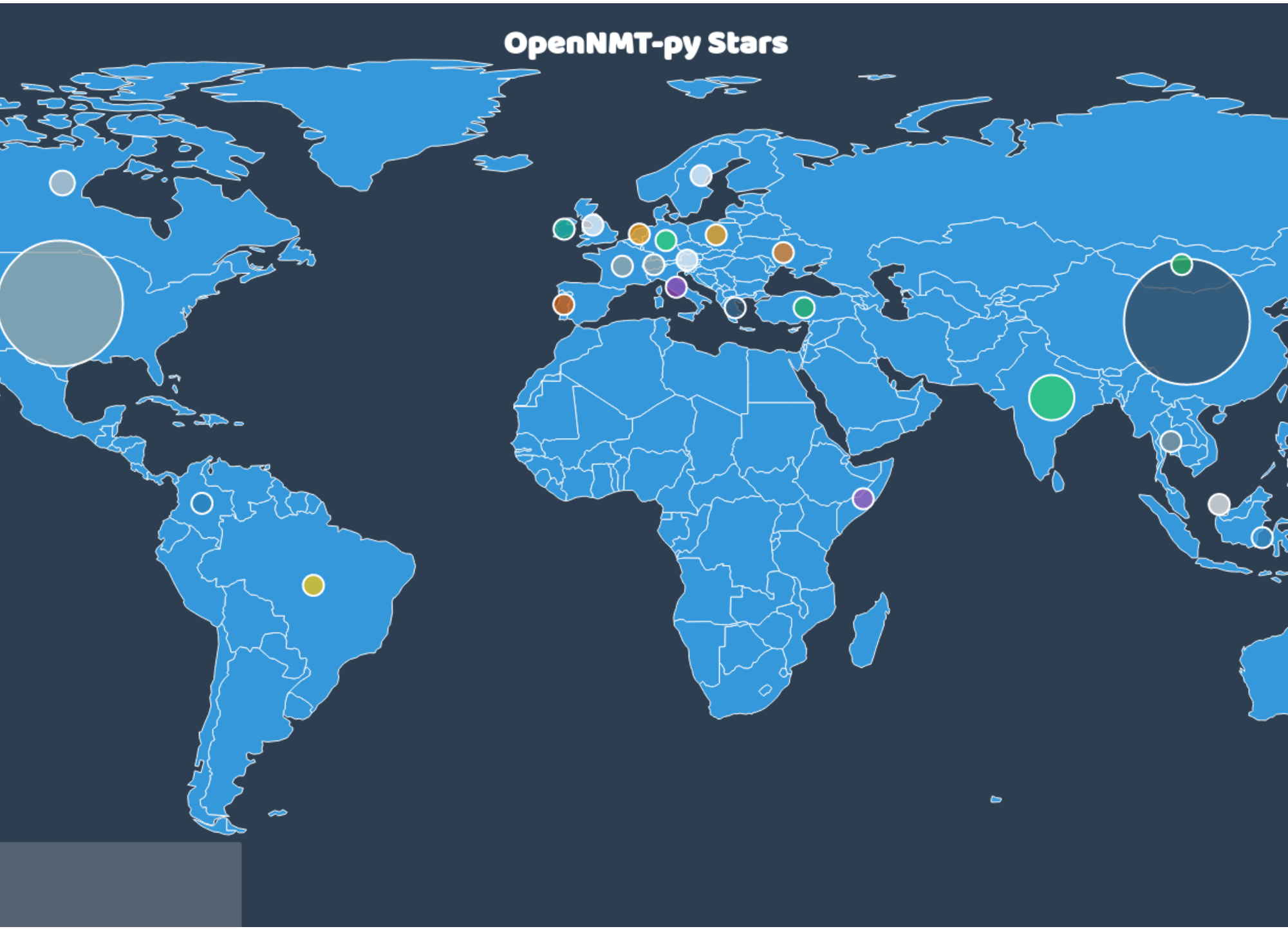
OpenNMT is an open source (MIT) initiative for neural machine translation and neural sequence modeling.



Since its launch in December 2016, OpenNMT has become a collection of implementations targeting both academia and industry. The systems are designed to be simple to use and easy to extend, while maintaining efficiency and state-of-the-art accuracy.

OpenNMT Statistics

- Almost 4,000 total GitHub stars and 1,000 forks.
- 3,500 code commits.
- 112 committers from across the world
- 3 major code bases (Lua, PyTorch, TensorFlow)
- Top 30 AI Project on GitHub
- Used by more that 60 research papers.
- Best Demo Runner-Up: Association of Computation Linguistics


OpenNMT-py Stars

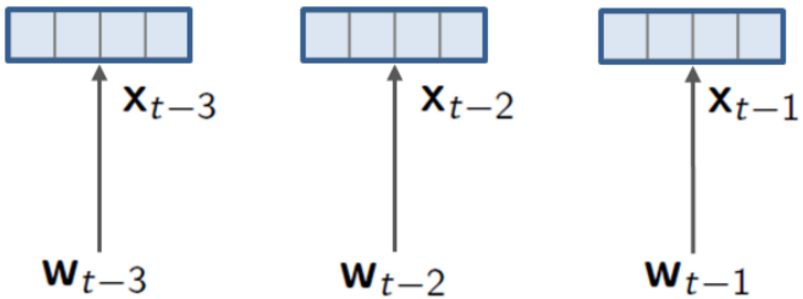
Today: A Worldwide Community

Neural Machine Translation: What is it?

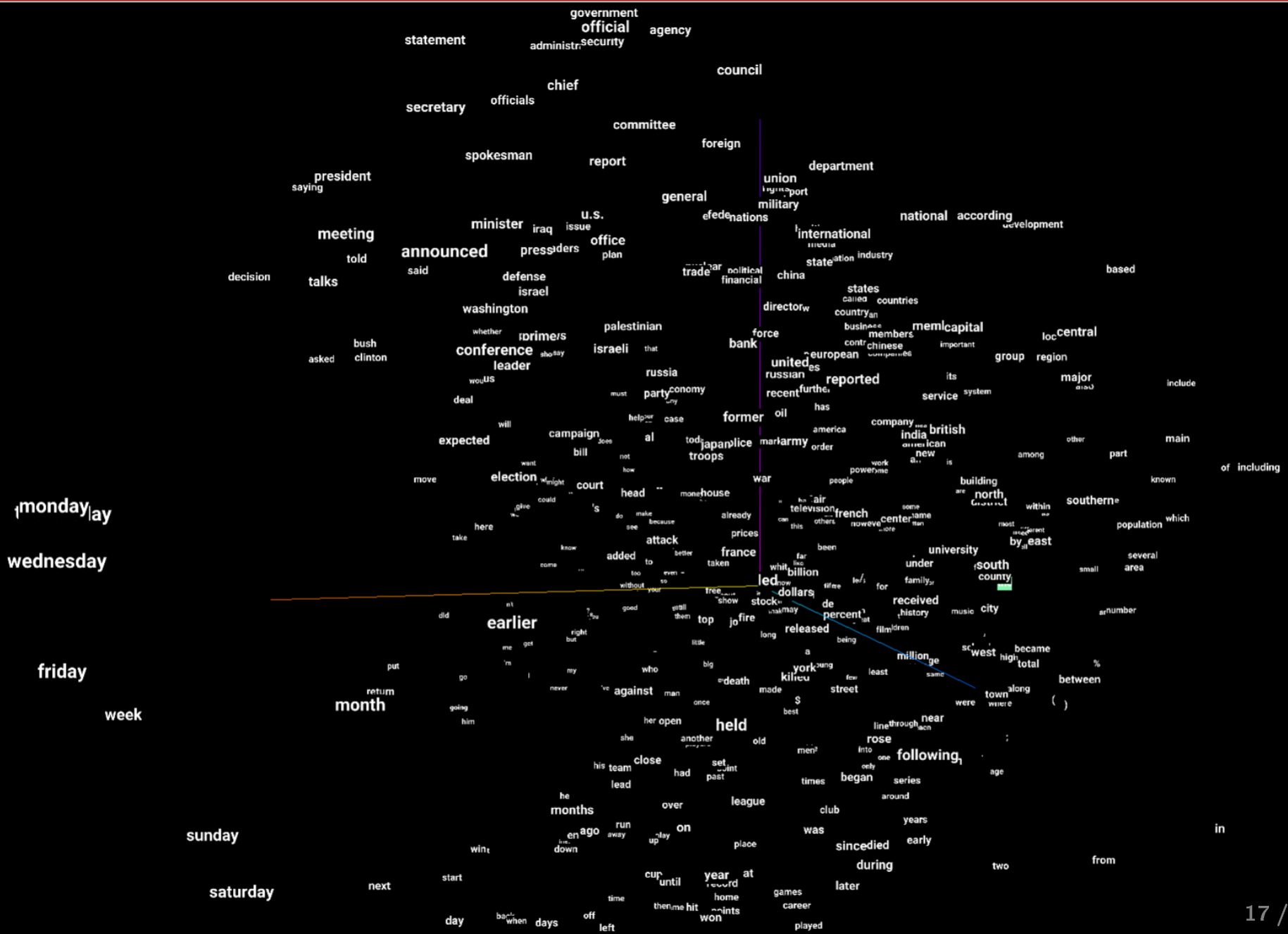
Modeling a Single Language

- Simple idea and diagrams, but requires data and parameters to scale.
- Languages have hundreds of thousands of words, and words are combined in an every possible way.
- Any sentence may have many different possible translations.

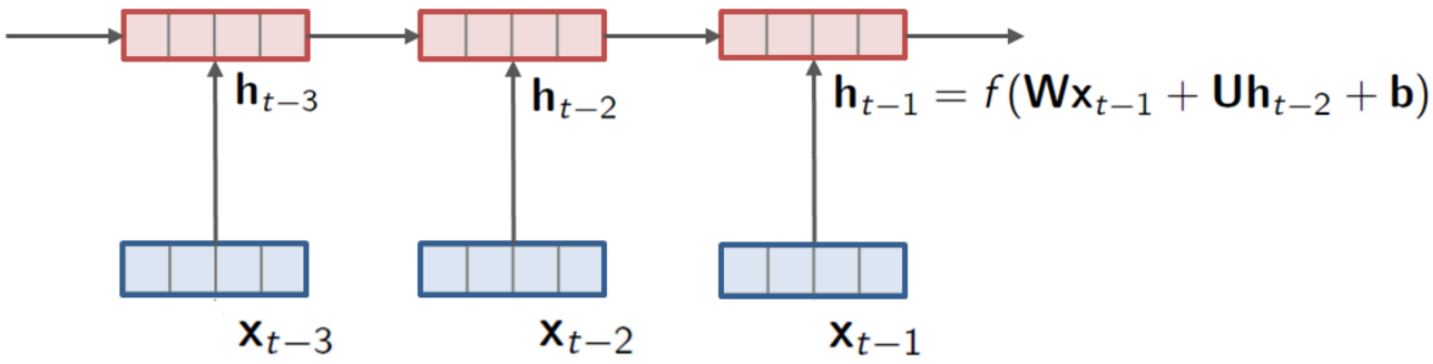
Step 1: Word Embeddings



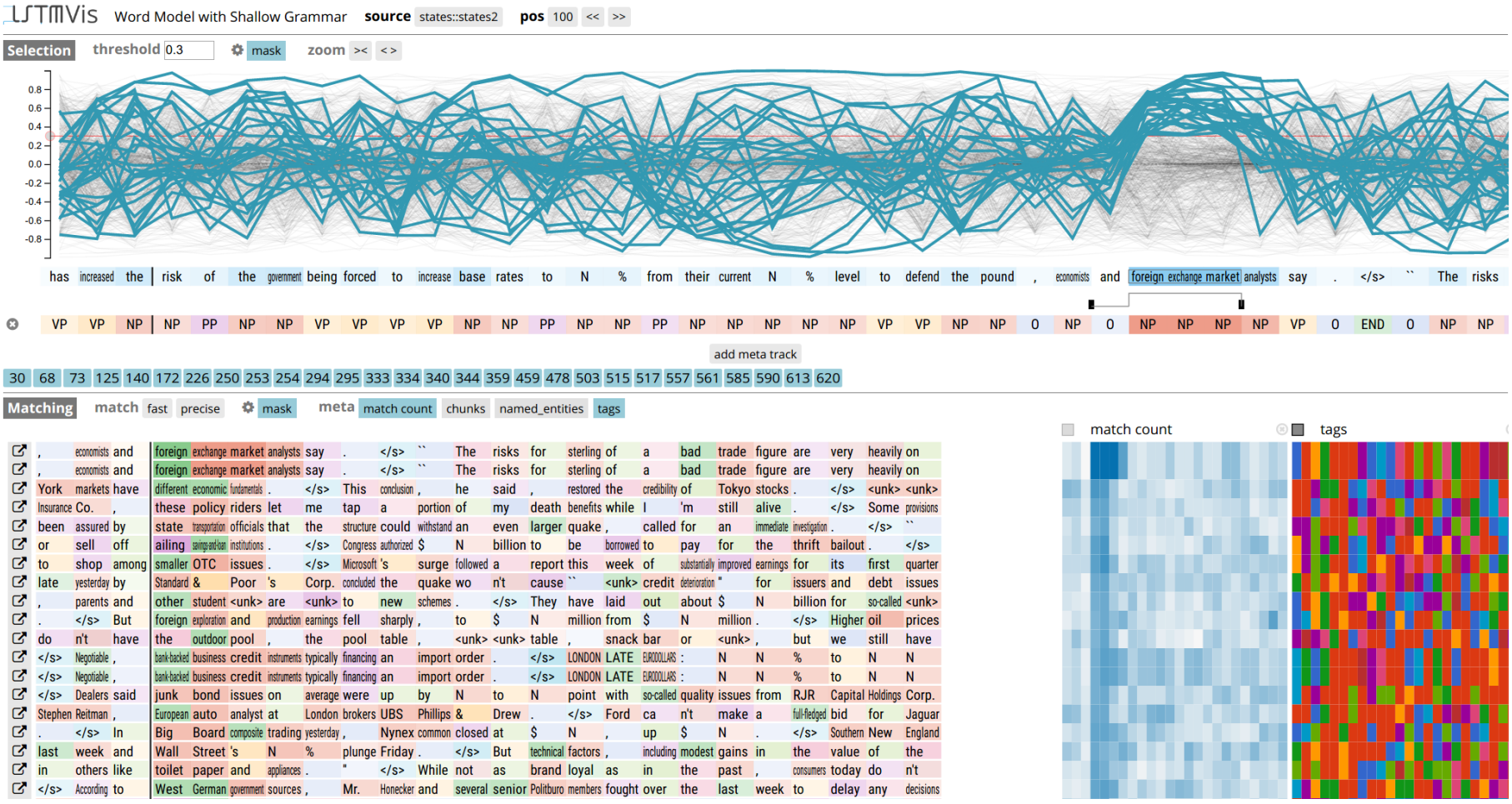
Step 1: Visualizing Word Representation (Tensorboard)



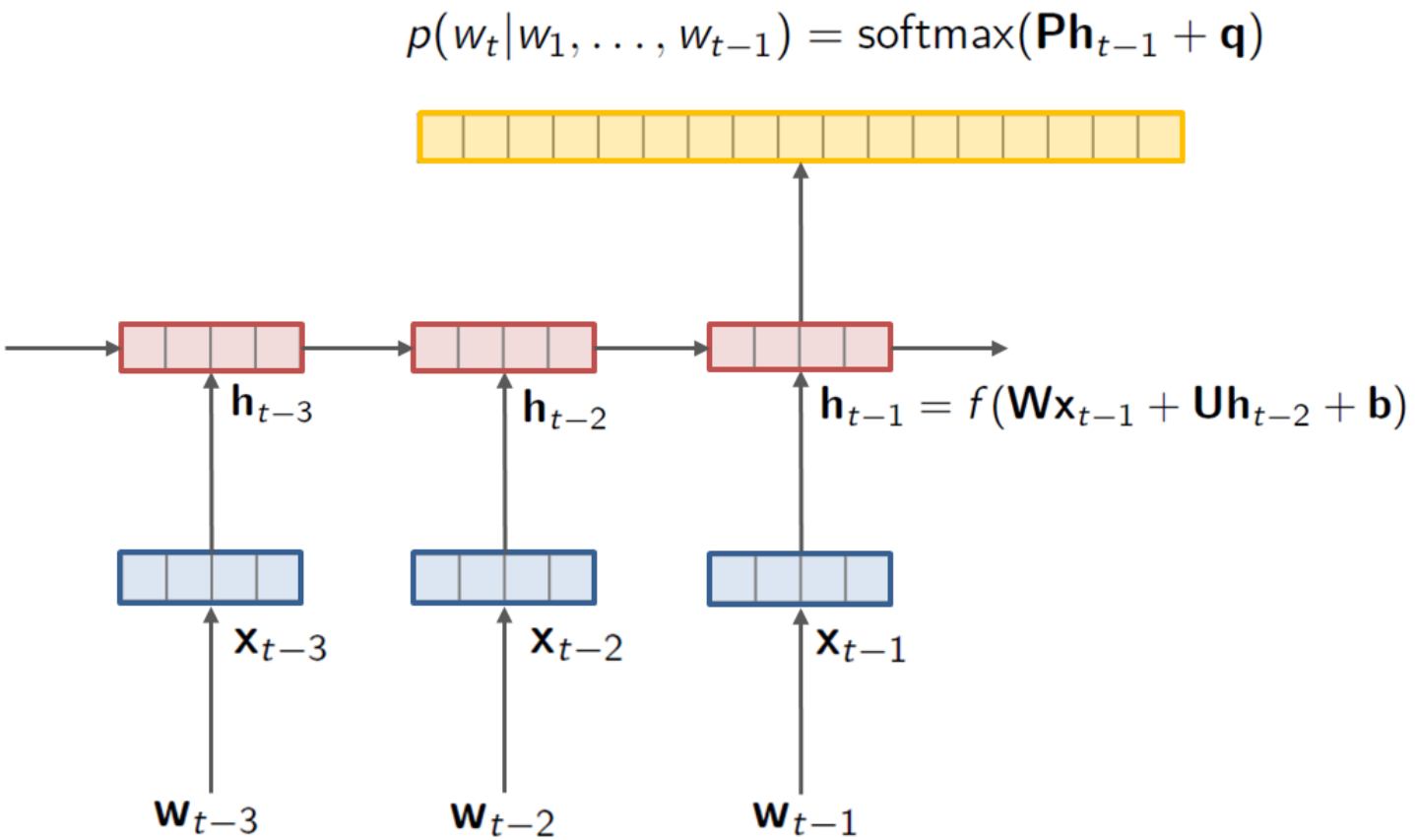
Step 2: Recurrent Neural Network



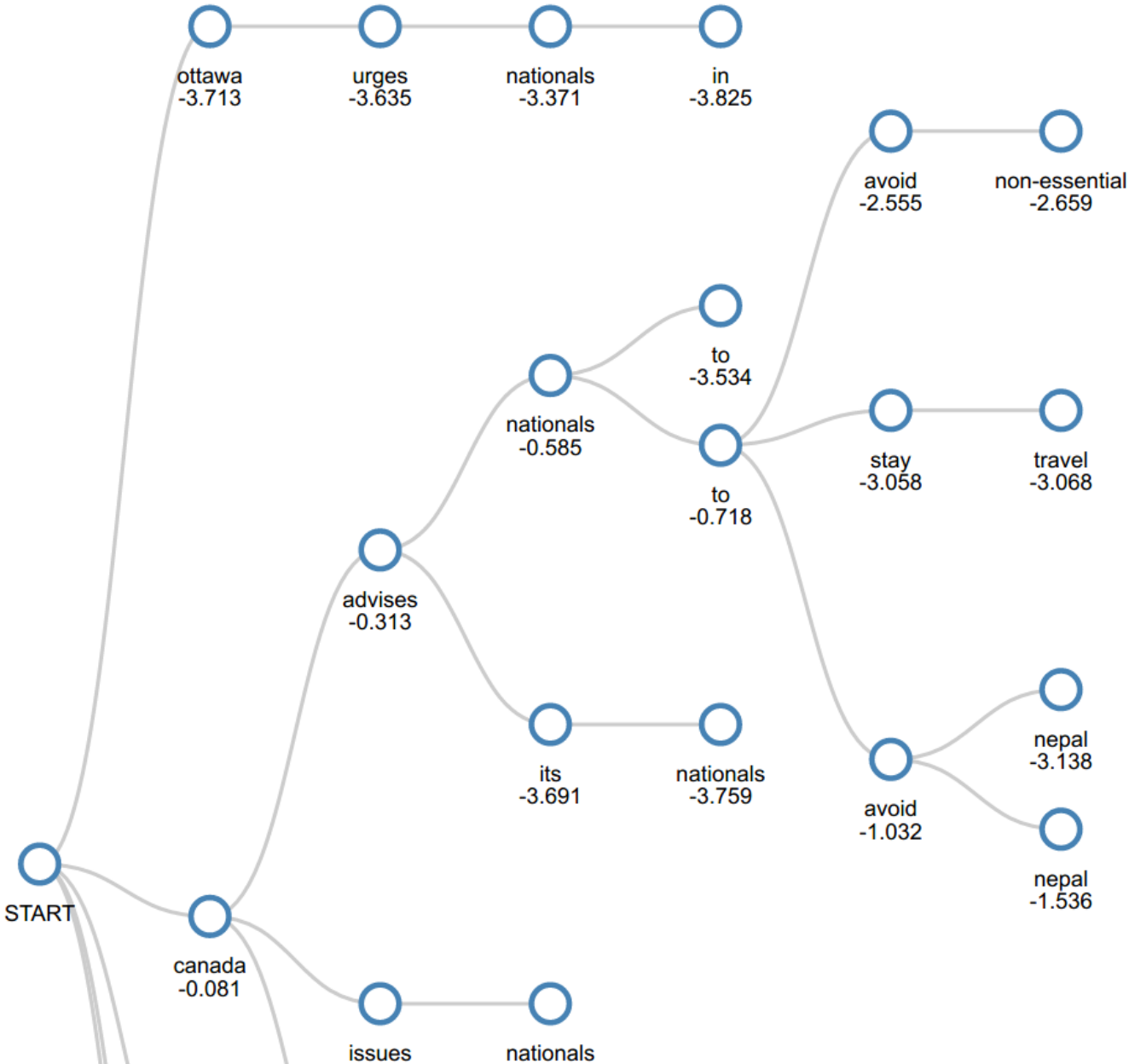
Step 2: Visualizing Sentence Representation



Step 3: Next Word Prediction



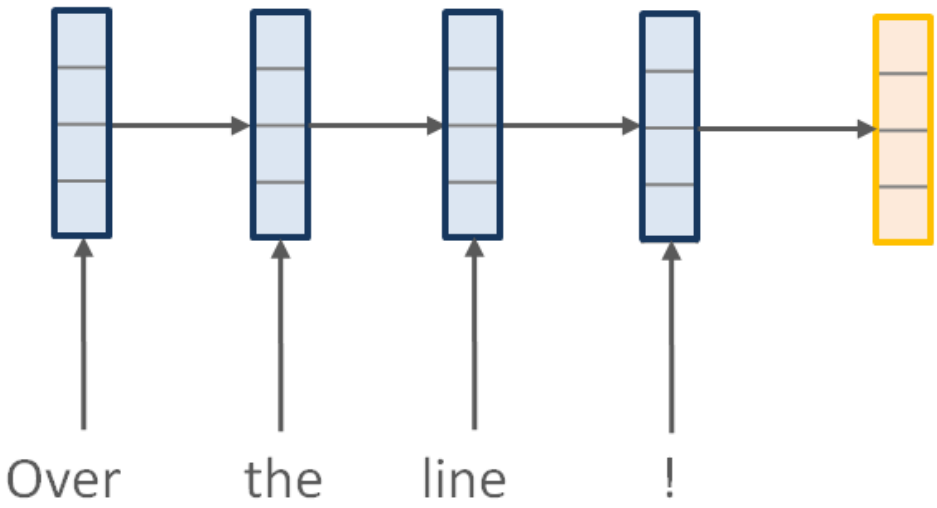
Step 3: Visualizing Word Prediction (NMT visualizer)



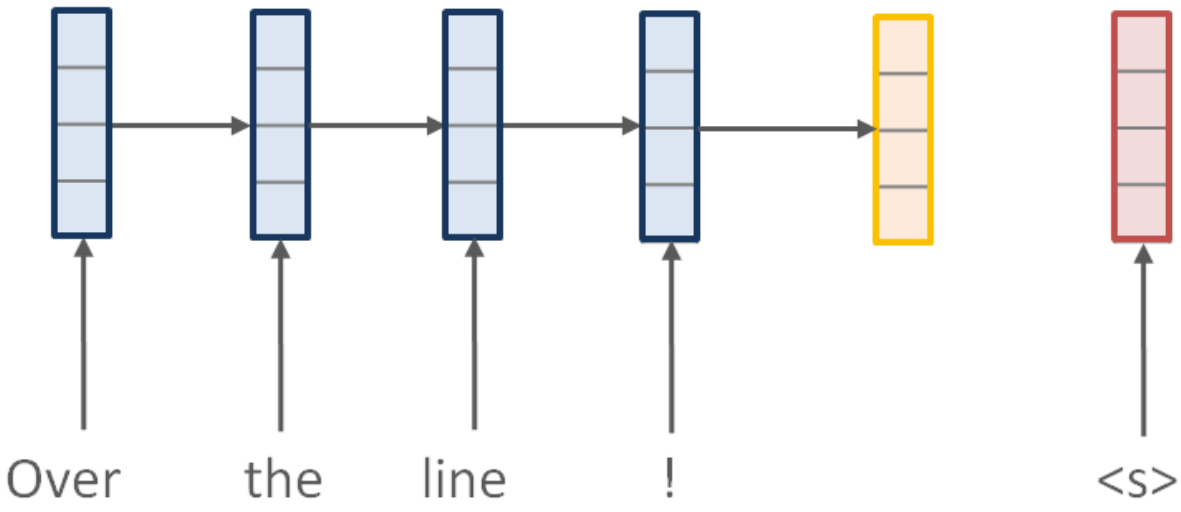
Example: Neural Machine Translation (Sutskever et al., 2014)

Over the line !

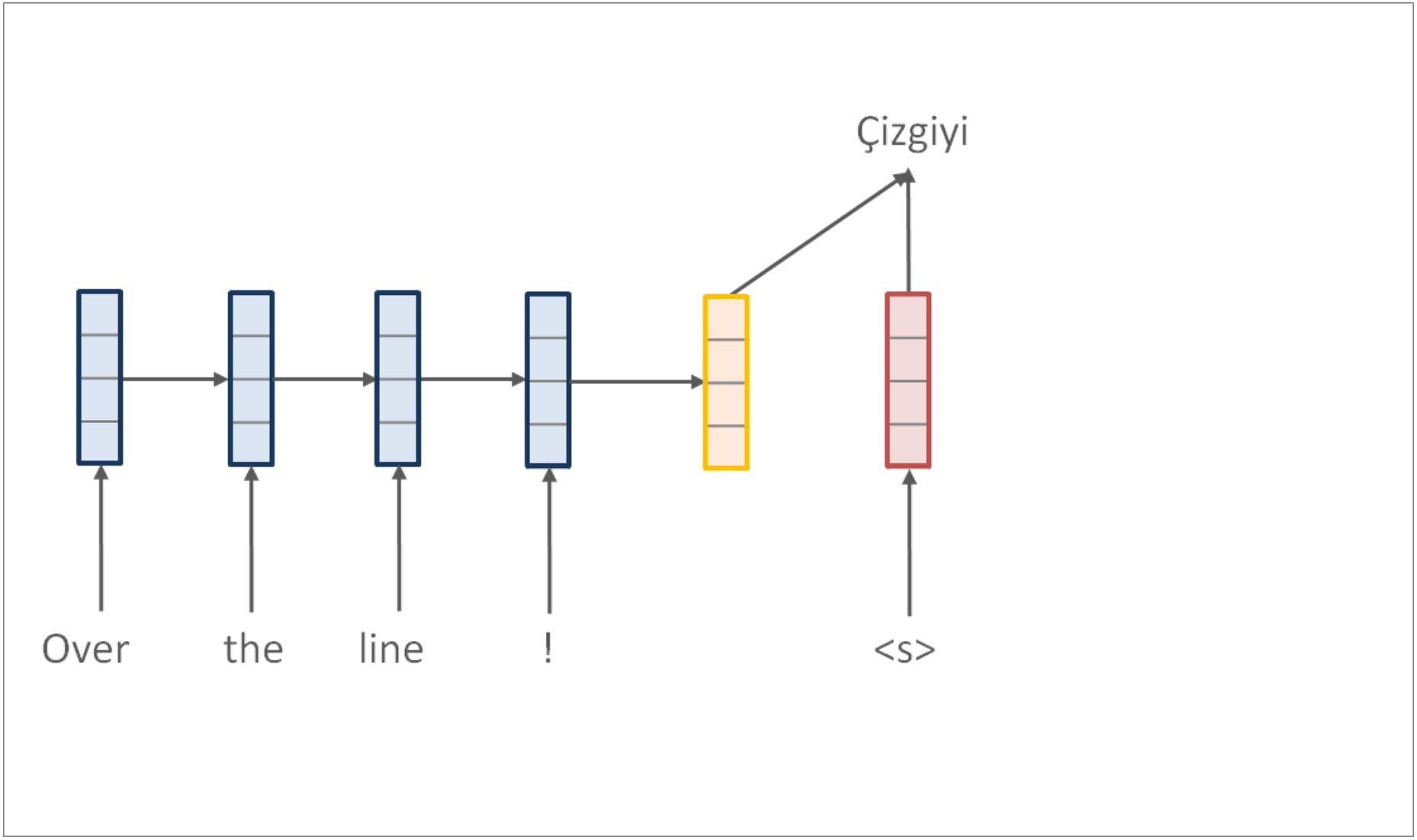
Example: Neural Machine Translation (Sutskever et al., 2014)



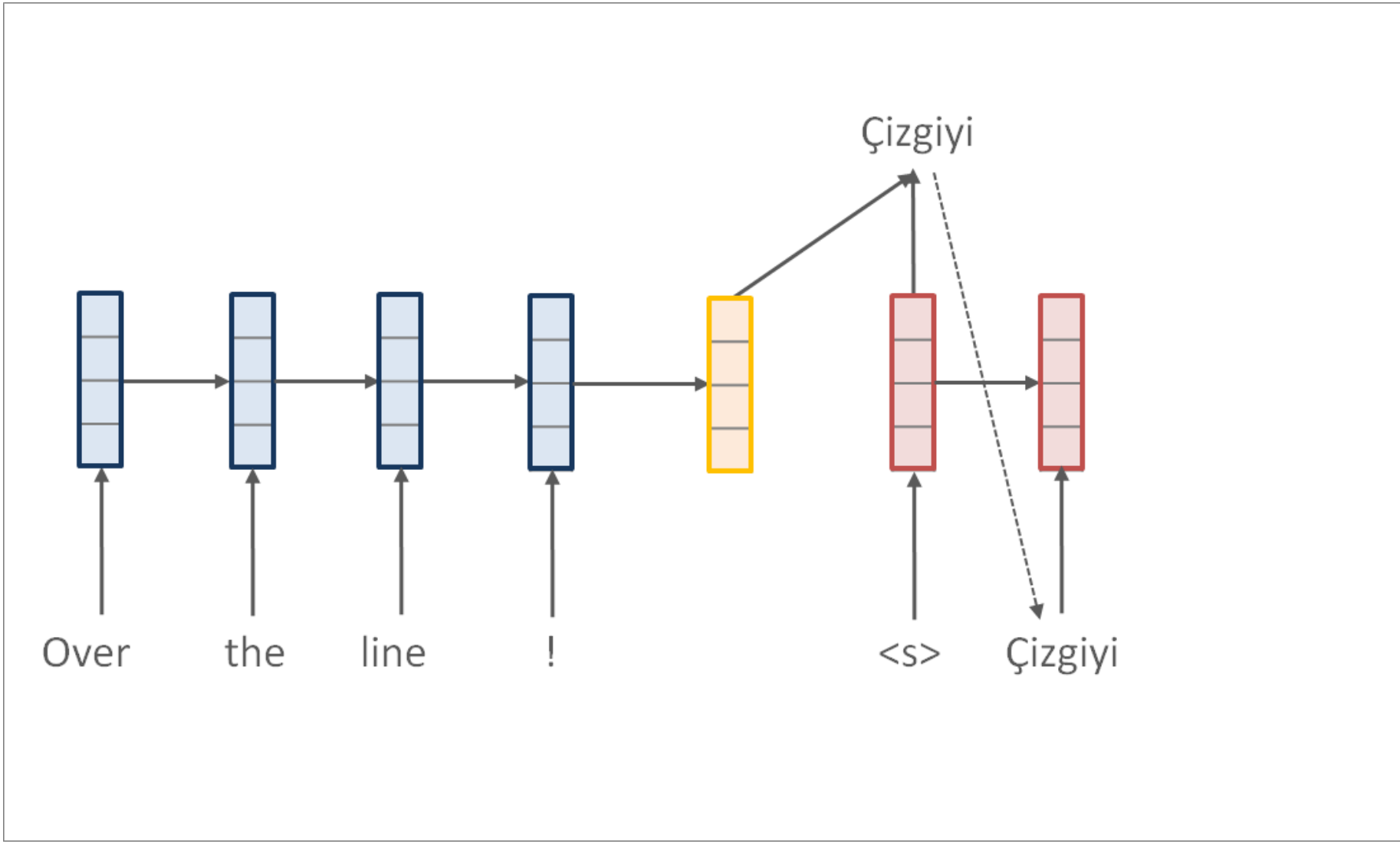
Example: Neural Machine Translation (Sutskever et al., 2014)



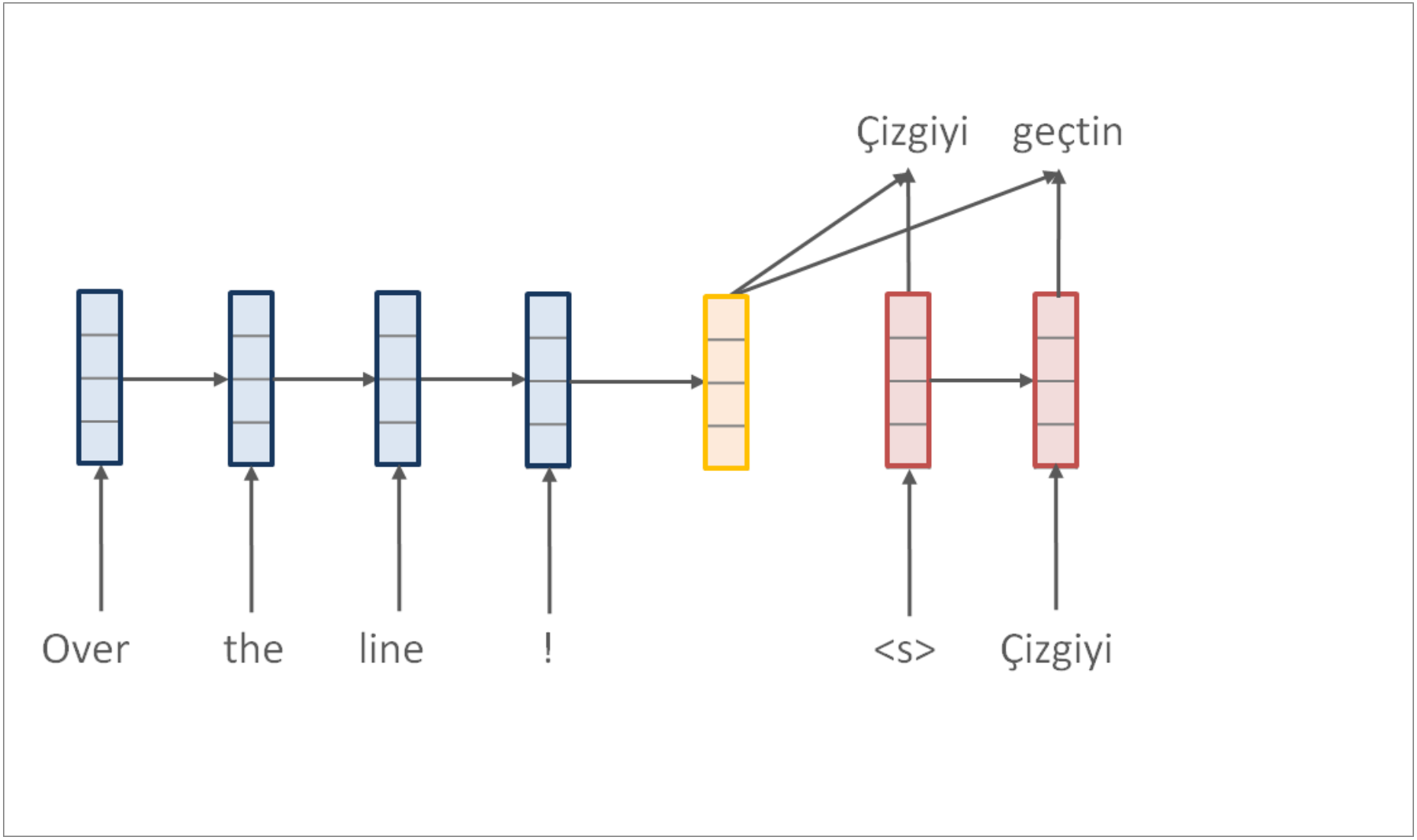
Example: Neural Machine Translation (Sutskever et al., 2014)



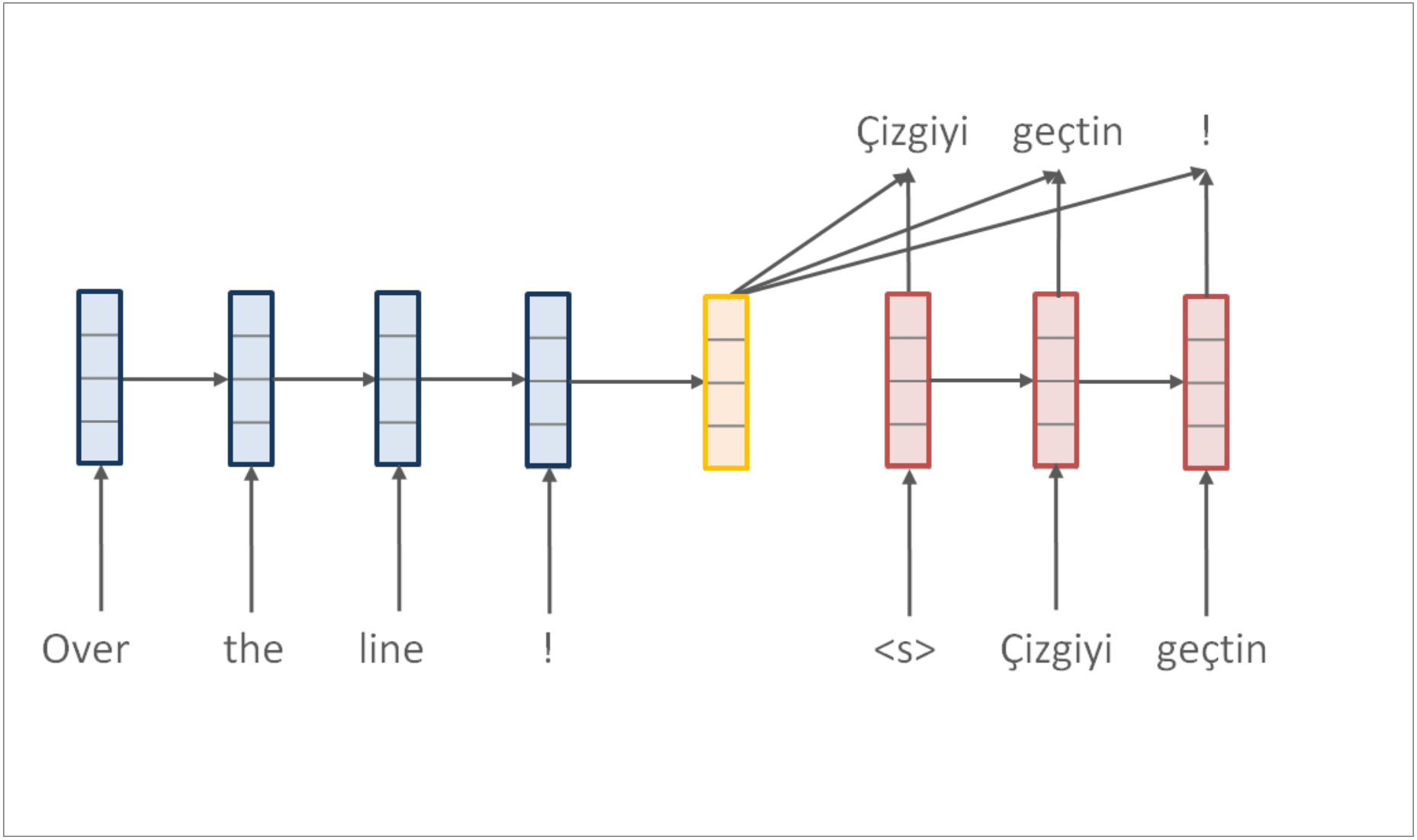
Example: Neural Machine Translation (Sutskever et al., 2014)



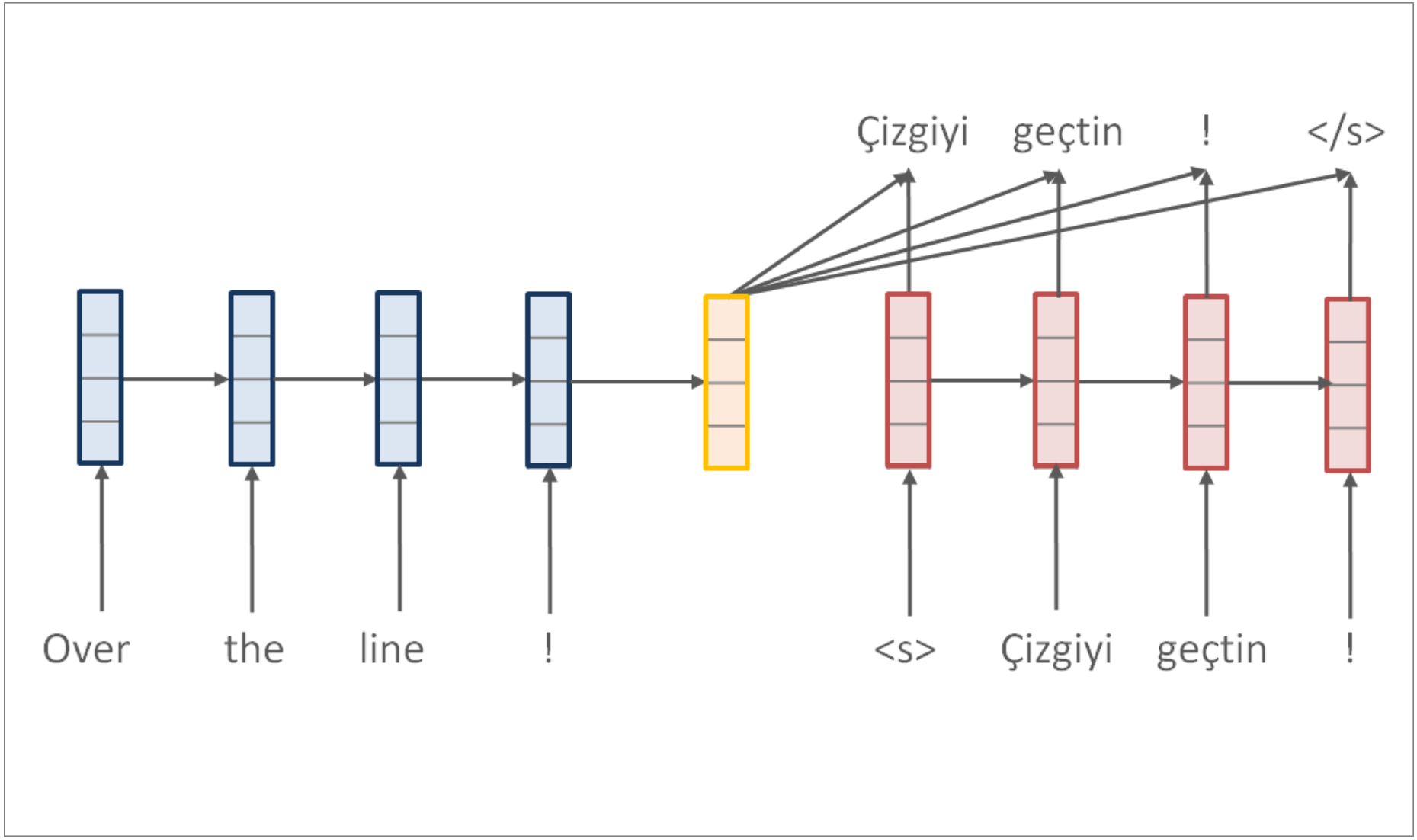
Example: Neural Machine Translation (Sutskever et al., 2014)



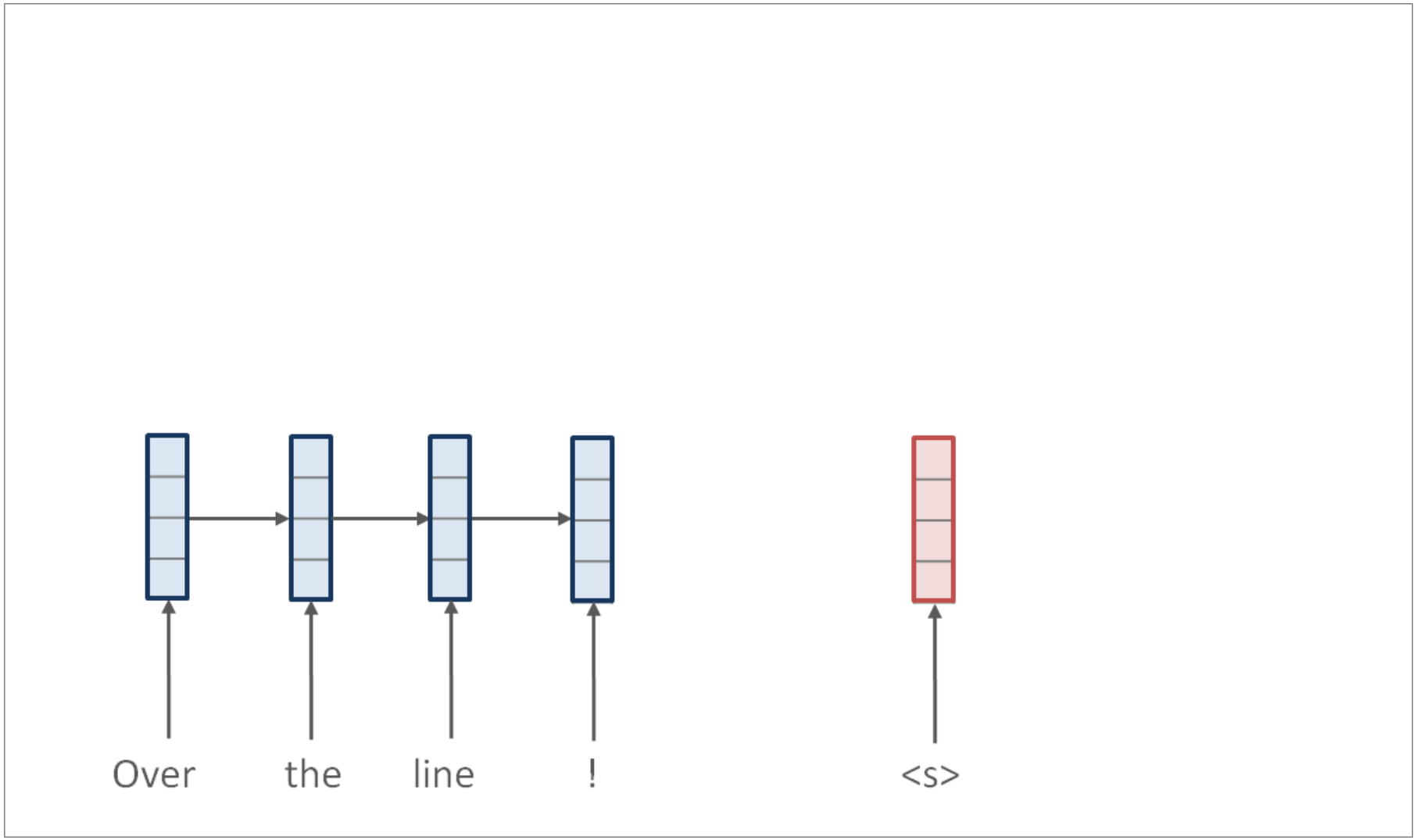
Example: Neural Machine Translation (Sutskever et al., 2014)



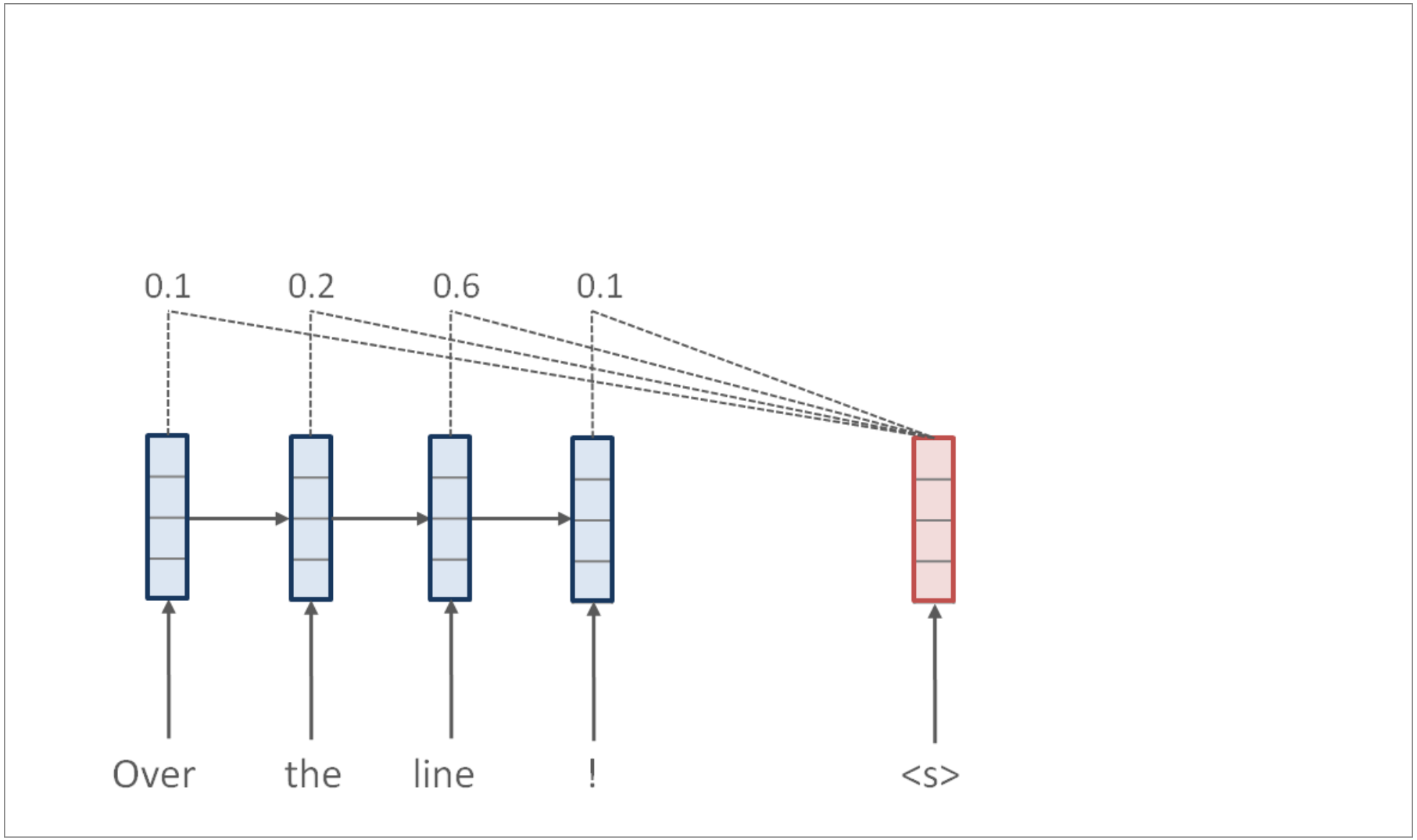
Example: Neural Machine Translation (Sutskever et al., 2014)



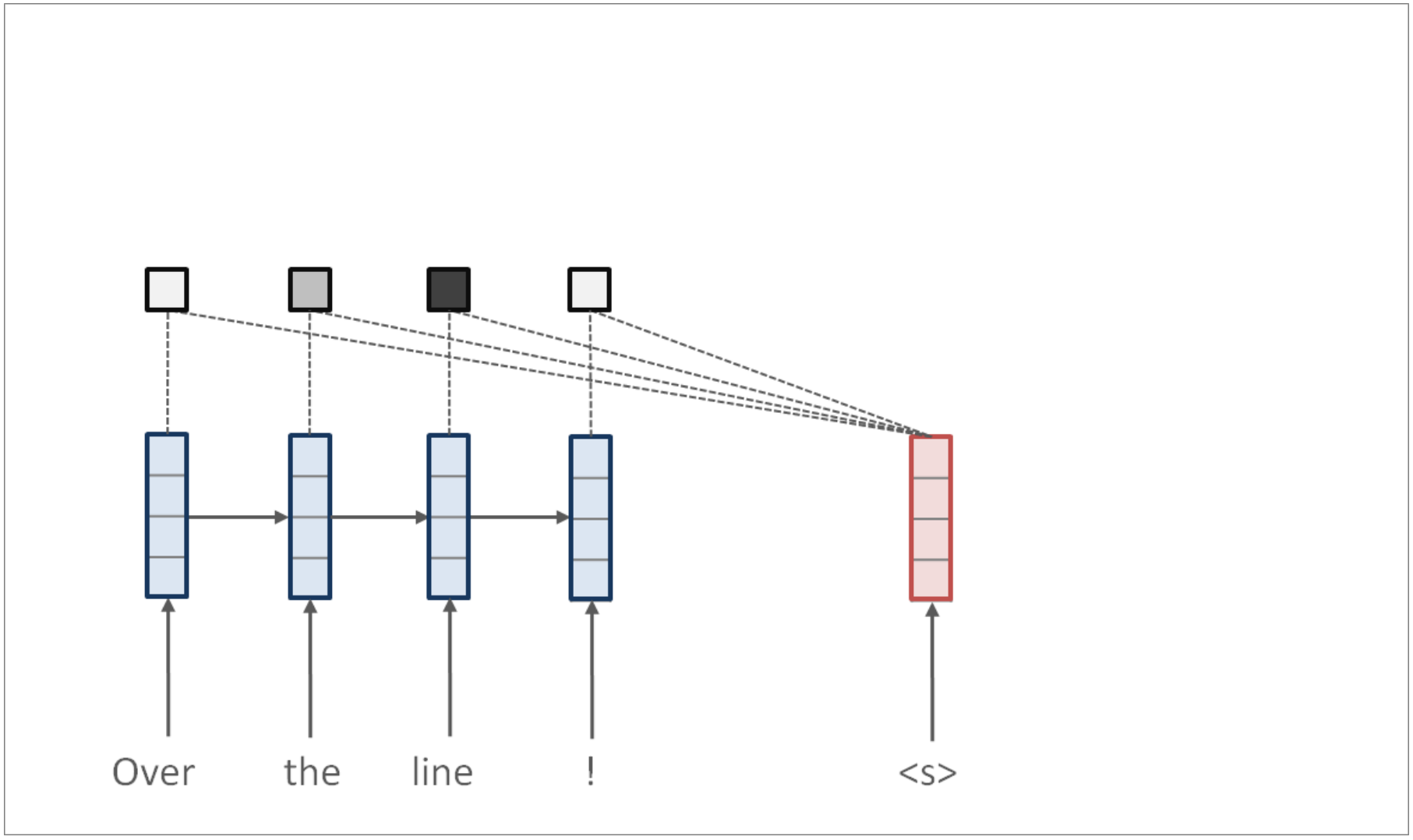
Attention-based Neural Machine Translation (Bahdanau et al., 2015)



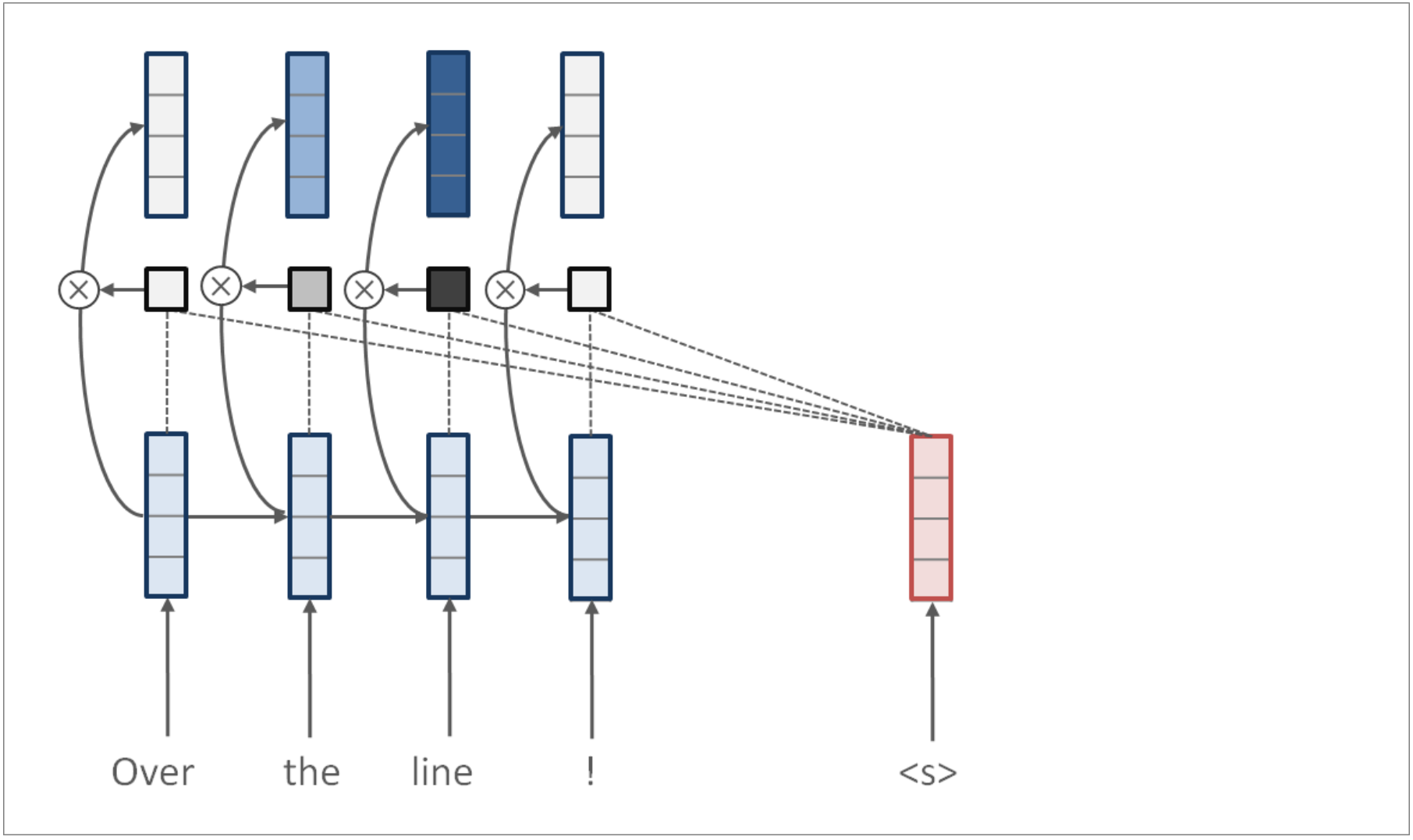
Attention-based Neural Machine Translation (Bahdanau et al., 2015)



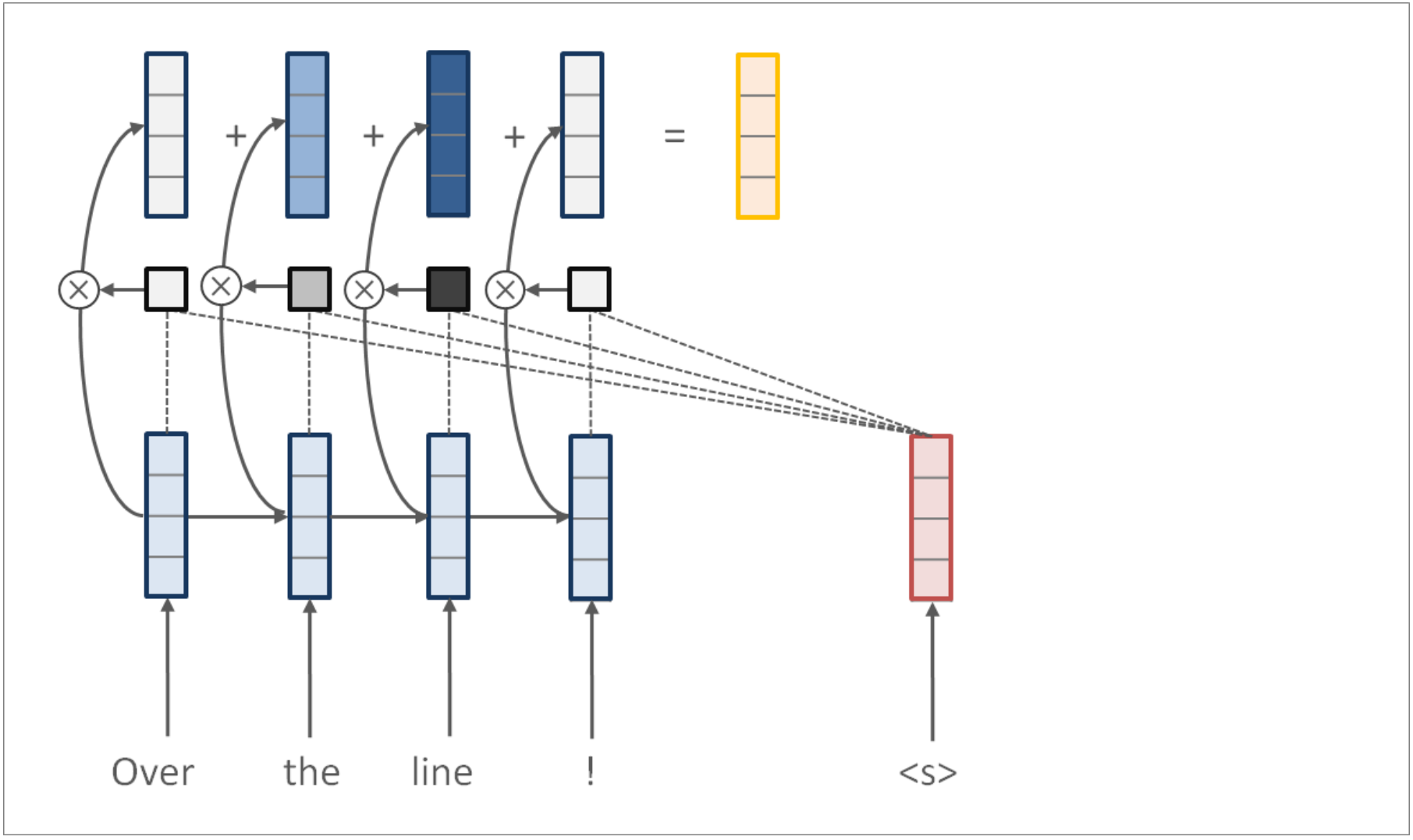
Attention-based Neural Machine Translation (Bahdanau et al., 2015)



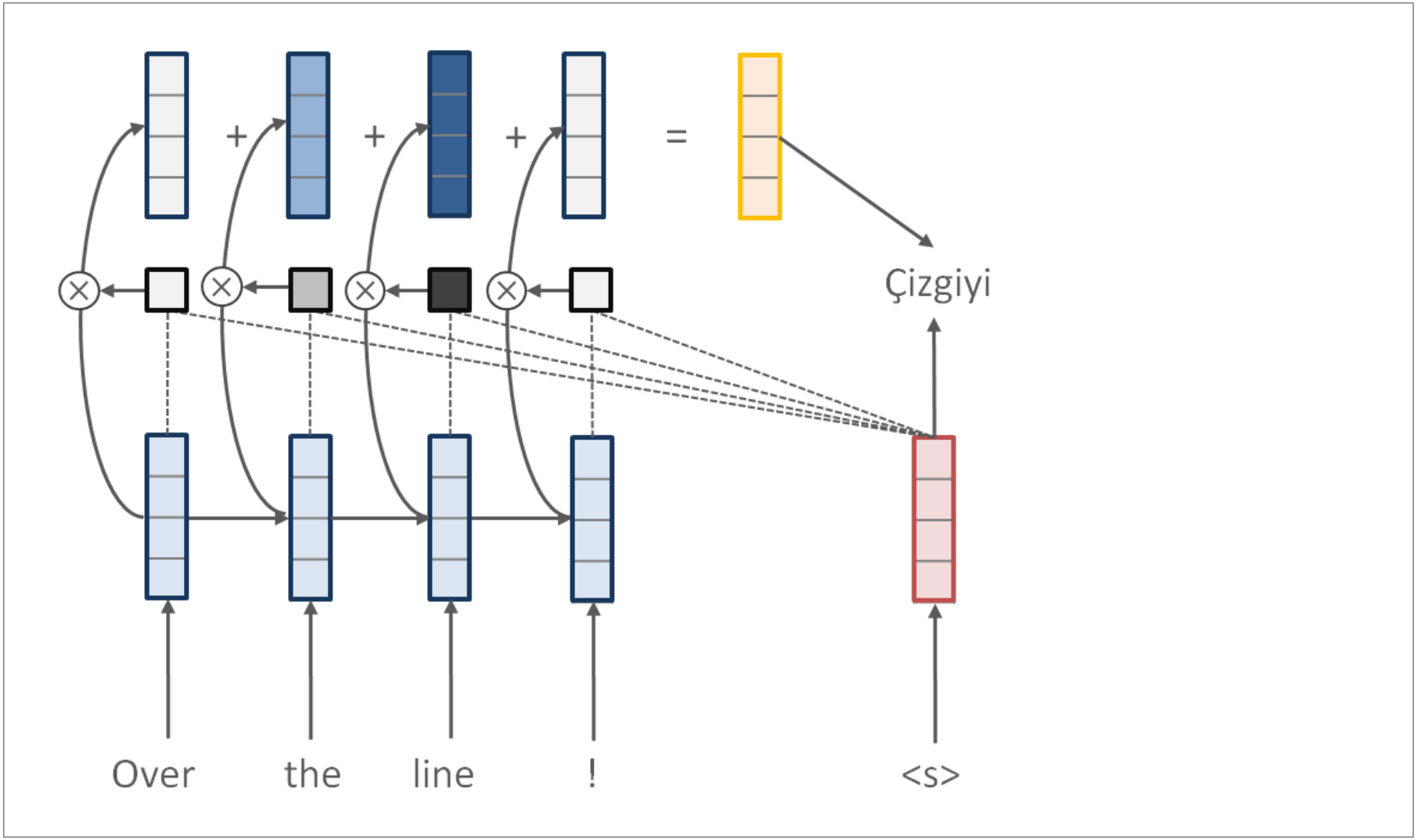
Attention-based Neural Machine Translation (Bahdanau et al., 2015)



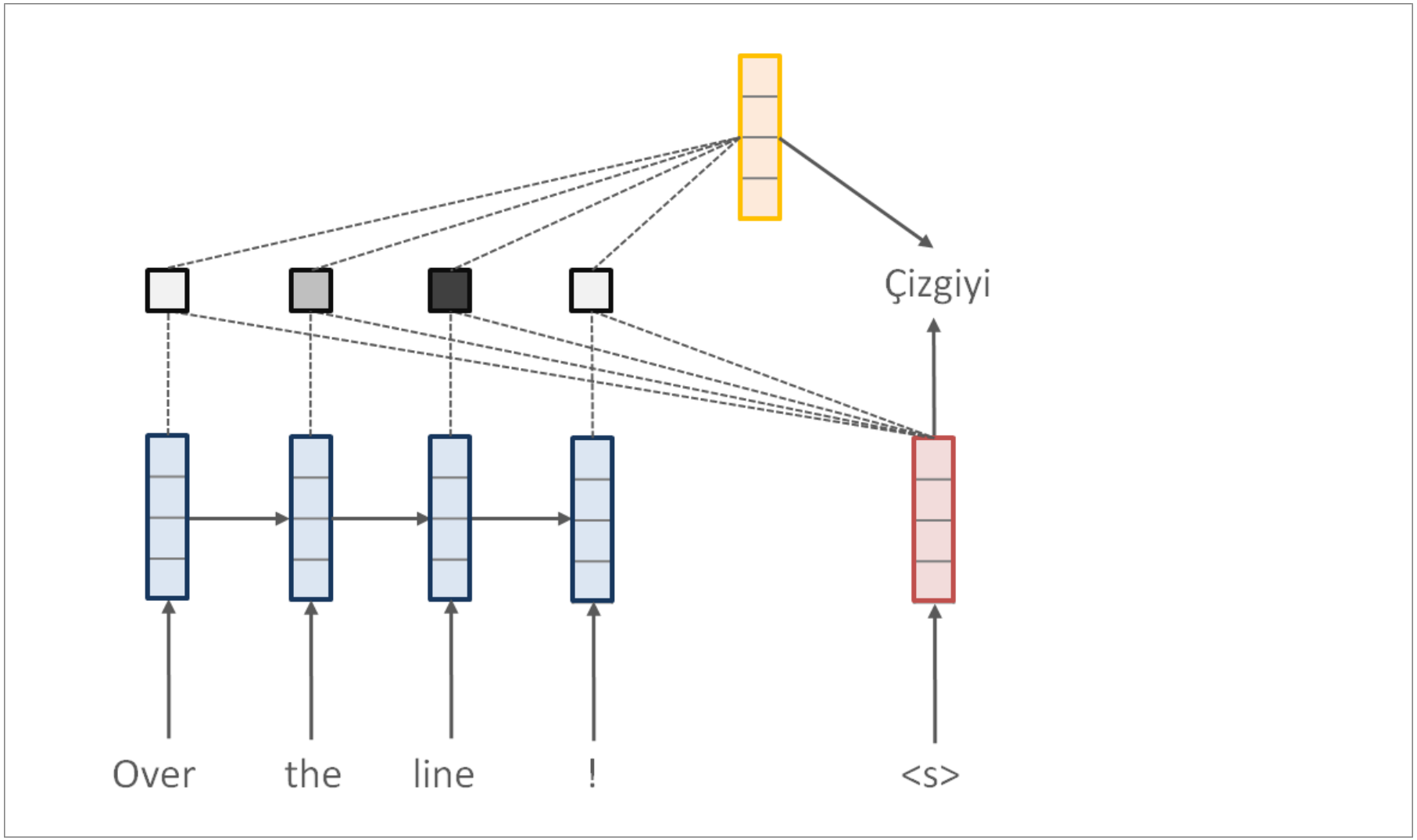
Attention-based Neural Machine Translation (Bahdanau et al., 2015)



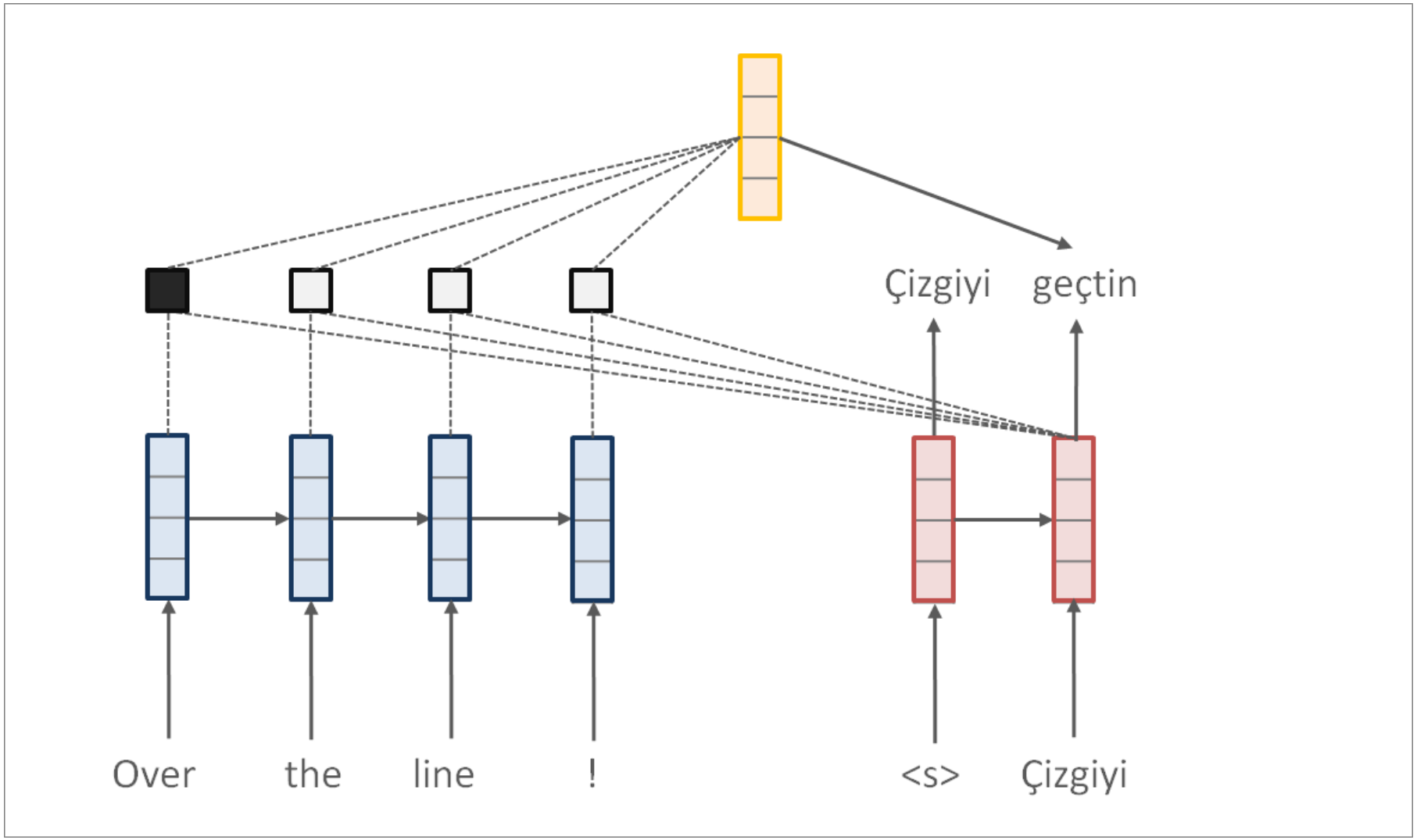
Attention-based Neural Machine Translation (Bahdanau et al., 2015)



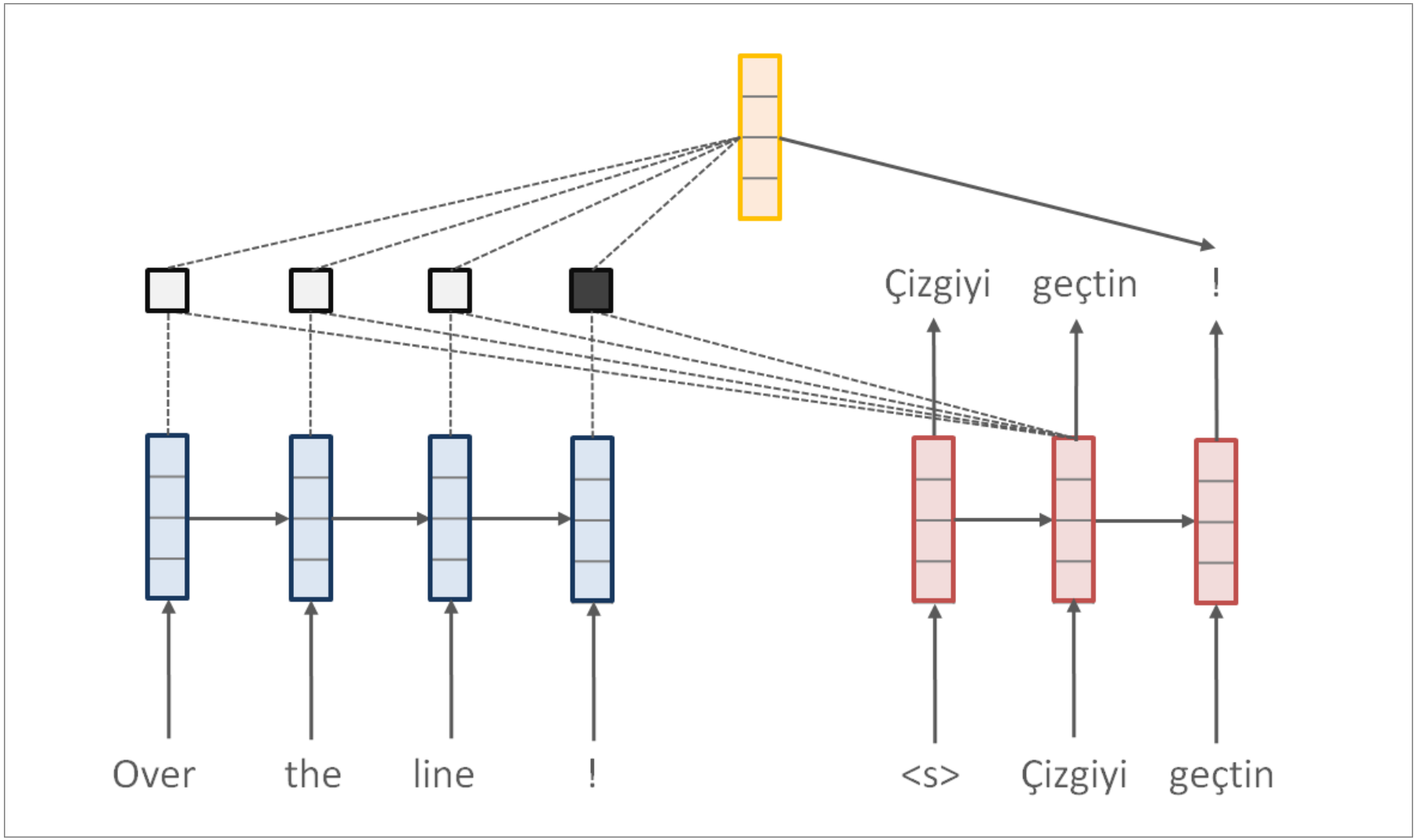
Attention-based Neural Machine Translation (Bahdanau et al., 2015)



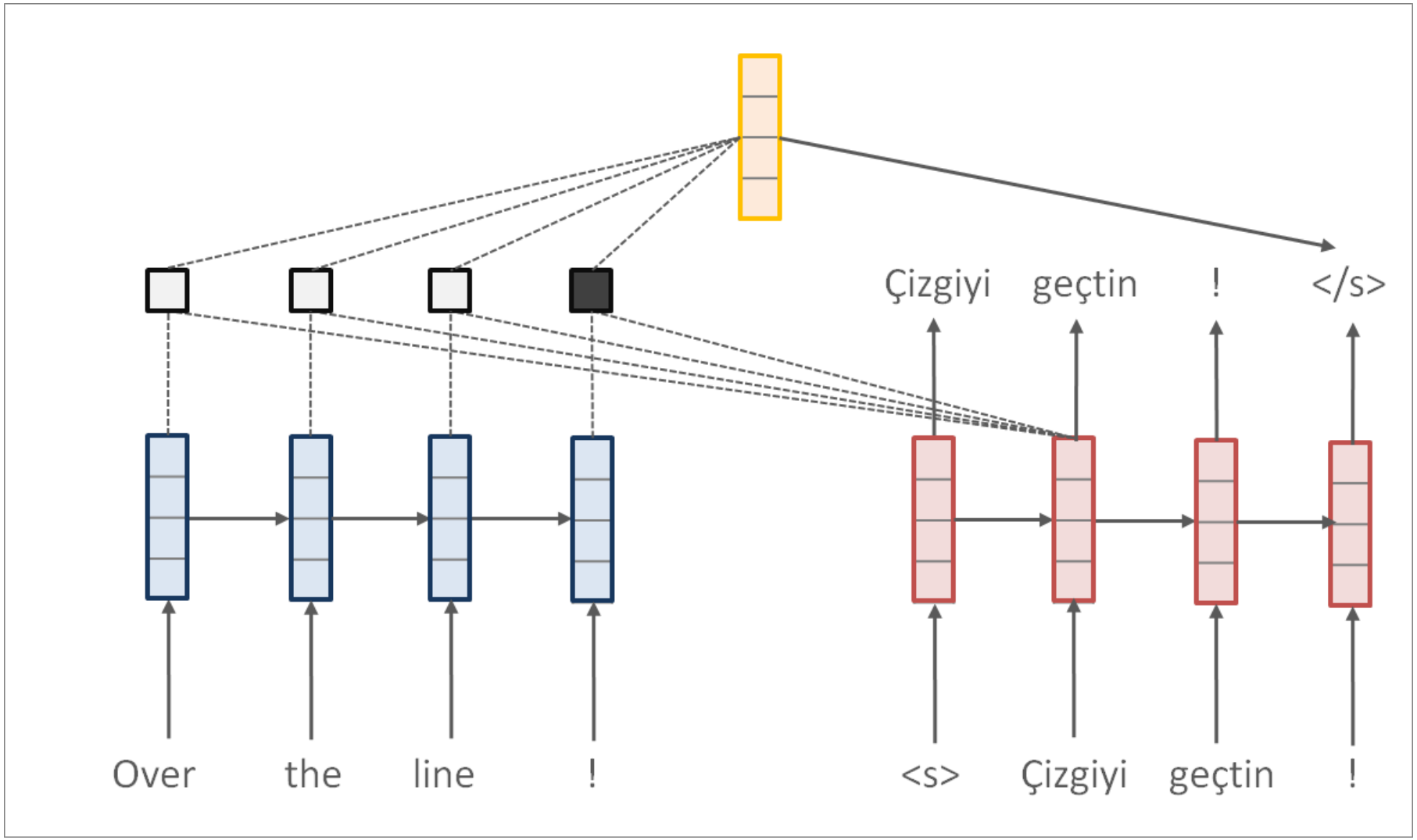
Attention-based Neural Machine Translation (Bahdanau et al., 2015)



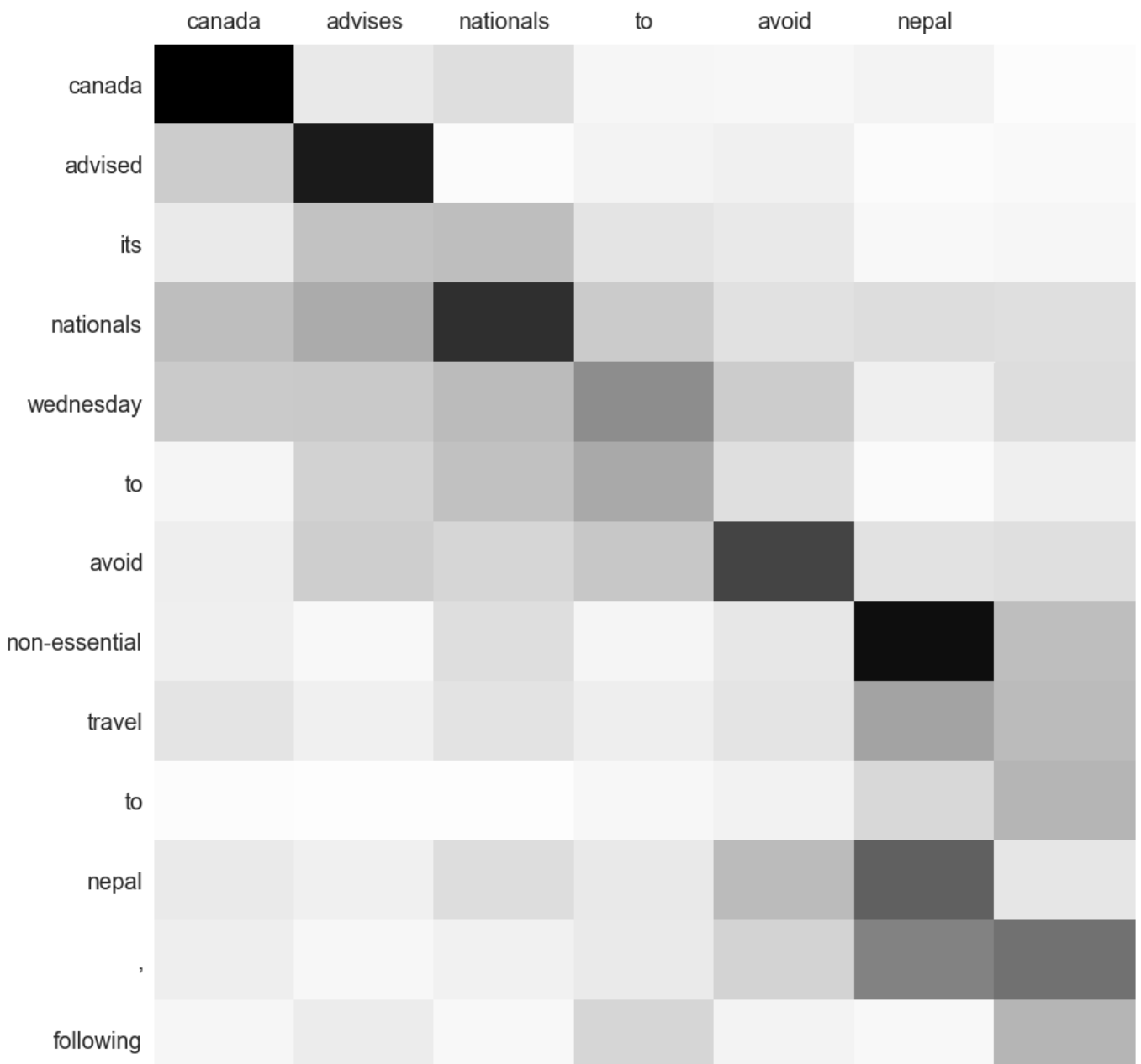
Attention-based Neural Machine Translation (Bahdanau et al., 2015)



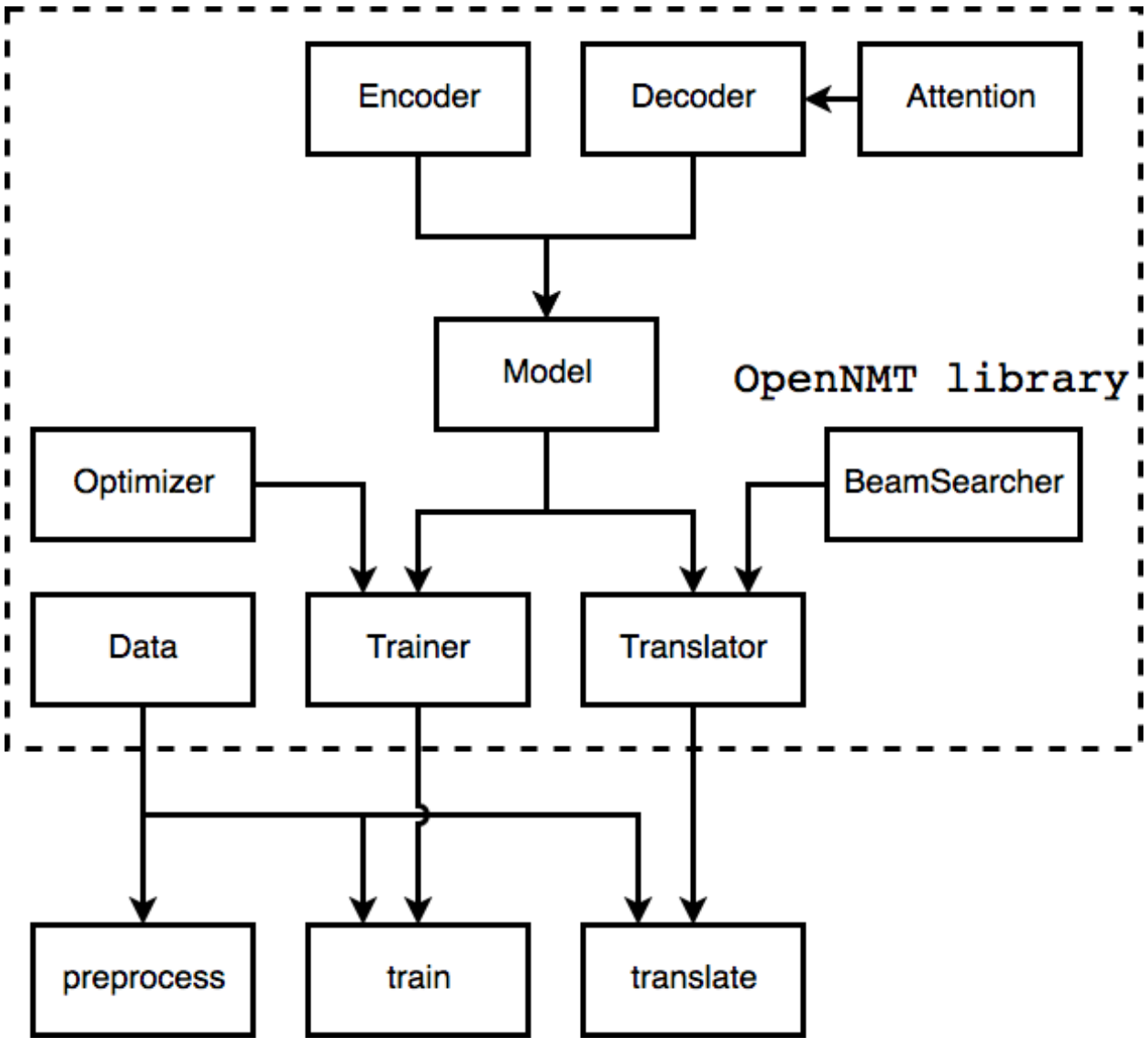
Attention-based Neural Machine Translation (Bahdanau et al., 2015)



Attention


canada
advises
nationals
to
avoid
nepal
canada
advised
its
nationals
wednesday
to
avoid
non-essential
travel
to
nepal
,
following

What is OpenNMT?

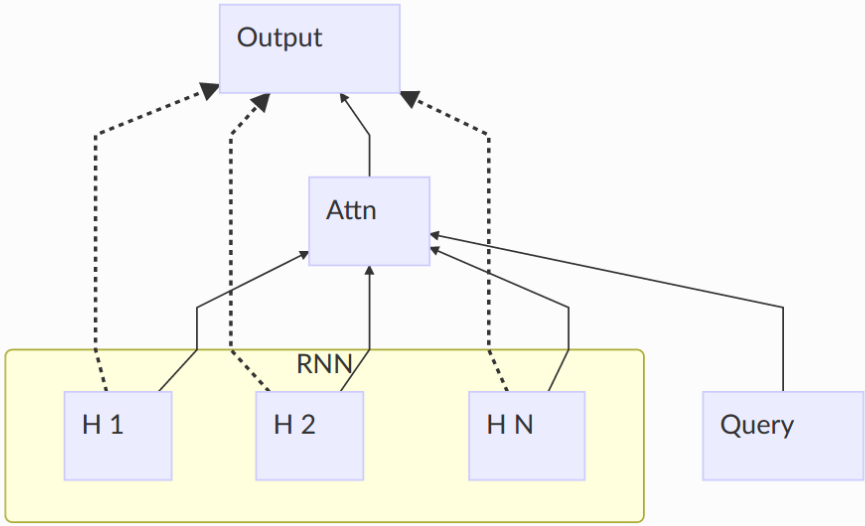


Education in OpenNMT

class `onmt.modules.GlobalAttention`(*dim, coverage=False, attn_type='dot'*)

Global attention takes a matrix and a query vector. It then computes a parameterized convex combination of the matrix based on the input query.

Constructs a unit mapping a query *q* of size *dim* and a source matrix *H* of size *n x dim*, to an output of size *dim*.



All models compute the output as $c = \sum_{j=1}^{SeqLength} a_j H_j$ where a_j is the softmax of a score function. Then then apply a projection layer to [q, c].

However they differ on how they compute the attention score.

- **Luong Attention (dot, general):**
  - dot: $score(H_j, q) = H_j^T q$
  - general: $score(H_j, q) = H_j^T W_a q$

What is next for the project?

E2E Text Generation:
Talk about Text (Summarization)

mexico city , mexico -lrb- cnn -rrb- – heavy rains and flooding have forced hundreds of thousands of people from homes in southern mexico 's state of tabasco over the past four days , with nearly as many trapped by the rising waters , state officials said thursday . officials say about 300,000 people are still trapped by the worst flooding in the region for 50 years . the grijalva river pushed over its banks through the state capital of villahermosa on thursday , forcing government workers to evacuate and leaving up to 80 percent of the city flooded , gov. andres granier 's office told cnn . about 700,000 people have seen their homes flooded , with about 300,000 of those still trapped there , granier 's office reported . one death had been blamed on the floods , which followed weeks of heavy rain in the largely swampy state . tabasco borders guatemala to the south and the gulf of mexico to the north . ...

E2E Text Generation:
Talk about Text (Summarization)

mexico city , mexico -lrb- cnn -rrb- – heavy rains and flooding have forced hundreds of thousands of people from homes in southern mexico 's state of tabasco over the past four days , with nearly as many trapped by the rising waters , state officials said thursday . officials say about 300,000 people are still trapped by the worst flooding in the region for 50 years . the grijalva river pushed over its banks through the state capital of villahermosa on thursday , forcing government workers to evacuate and leaving up to 80 percent of the city flooded , gov. andres granier 's office told cnn . about 700,000 people have seen their homes flooded , with about 300,000 of those still trapped there , granier 's office reported . one death had been blamed on the floods , which followed weeks of heavy rain in the largely swampy state . tabasco borders guatemala to the south and the gulf of mexico to the north . ...

tabasco and chiapas states hardest hit. authorities say 700,000 affected ...

E2E Generation Challenge:
Talk about the Environment (Multimodal)

E2E Generation Challenge:
Talk about the Environment (Multimodal)

Dow and S&P 500 close out week at all-time highs ...

E2E Text Generation:
Talk about Information (Generation)

TEAM	W	L	PTS	...
Heat	11	12	103	...
Hawks	7	15	95	...

E2E Text Generation:
Talk about Information (Generation)

TEAM	W	L	PTS	...
Heat	11	12	103	...
Hawks	7	15	95	...

The Atlanta Hawks defeated the Miami Heat, 103 - 95, at Philips Arena on Wednesday. Atlanta ...

2018: OpenNMT For Everything

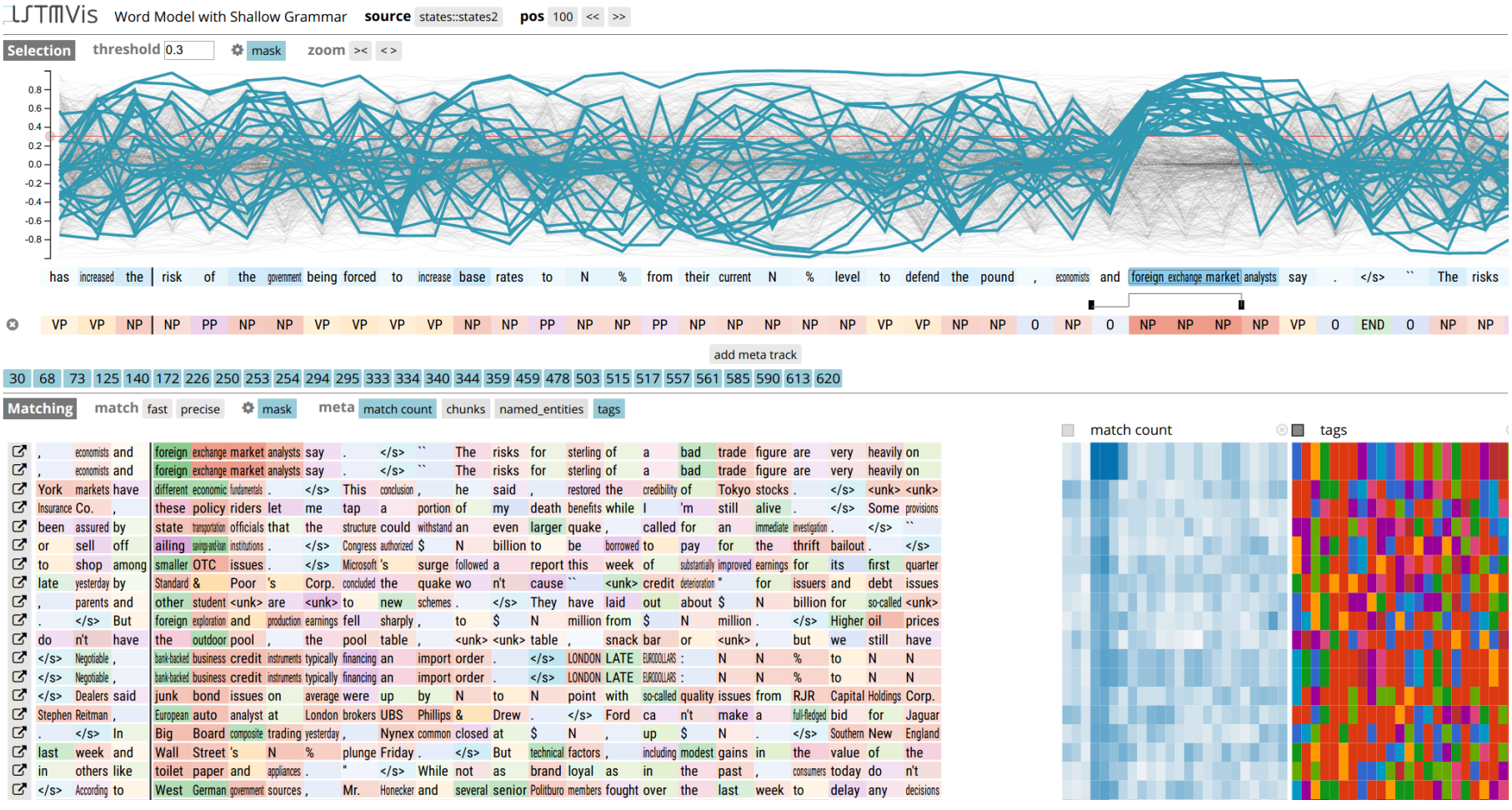
English Summarization

Who/When	Corpus Prep	Training Tool	Training Parameters	Server Details	Training Time/Memory	Translation Parameters	Scores	Model
2018/02/11 Baseline	Gigaword Standard	OpenNMT `d4ab35a`	2 layers, RNN 500, WE 500, input feed 20 epochs	Trained on 1 GPU TITAN X			Gigaword F-Score R1: 33.60 R2: 16.29 RL: 31.45	331MB here
2018/02/22 Baseline	Gigaword Standard	OpenNMT `338b3b1`	2 layers, RNN 500, WE 500, input feed, copy_attn, reuse_copy_attn 20 epochs	Trained on 1 GPU TITAN X		replace_unk	Gigaword F-Score R1: 35.51 R2: 17.35 RL: 33.17	331MB here

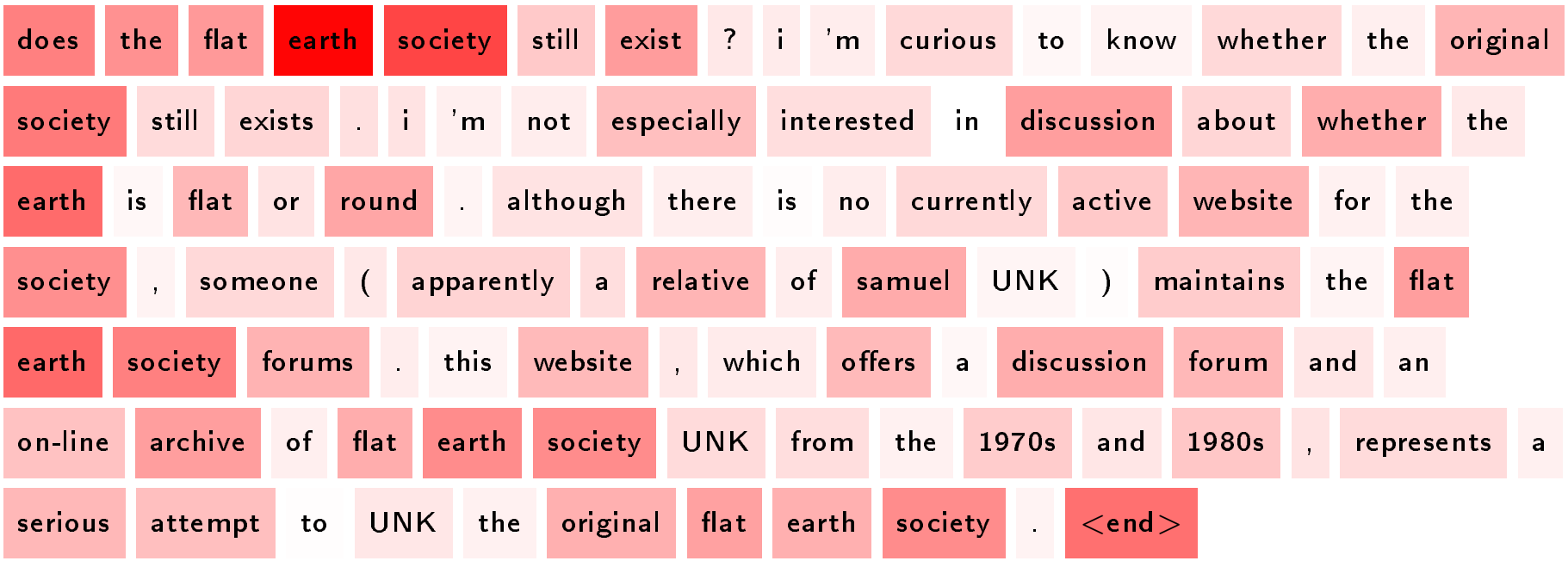
Dialog System

Who/When	Corpus Prep	Training Tool	Training Parameters	Server Details	Training Time/Memory	Translation Parameters	Scores	Model
2018/02/22 Baseline	Opensubtitles	OpenNMT `338b3b1`	2 layers, RNN 500, WE 500, input feed, dropout 0.2, global_attention mlp, start_decay_at 7 13 epochs	Trained on 1 GPU TITAN X			TBD	355MB here

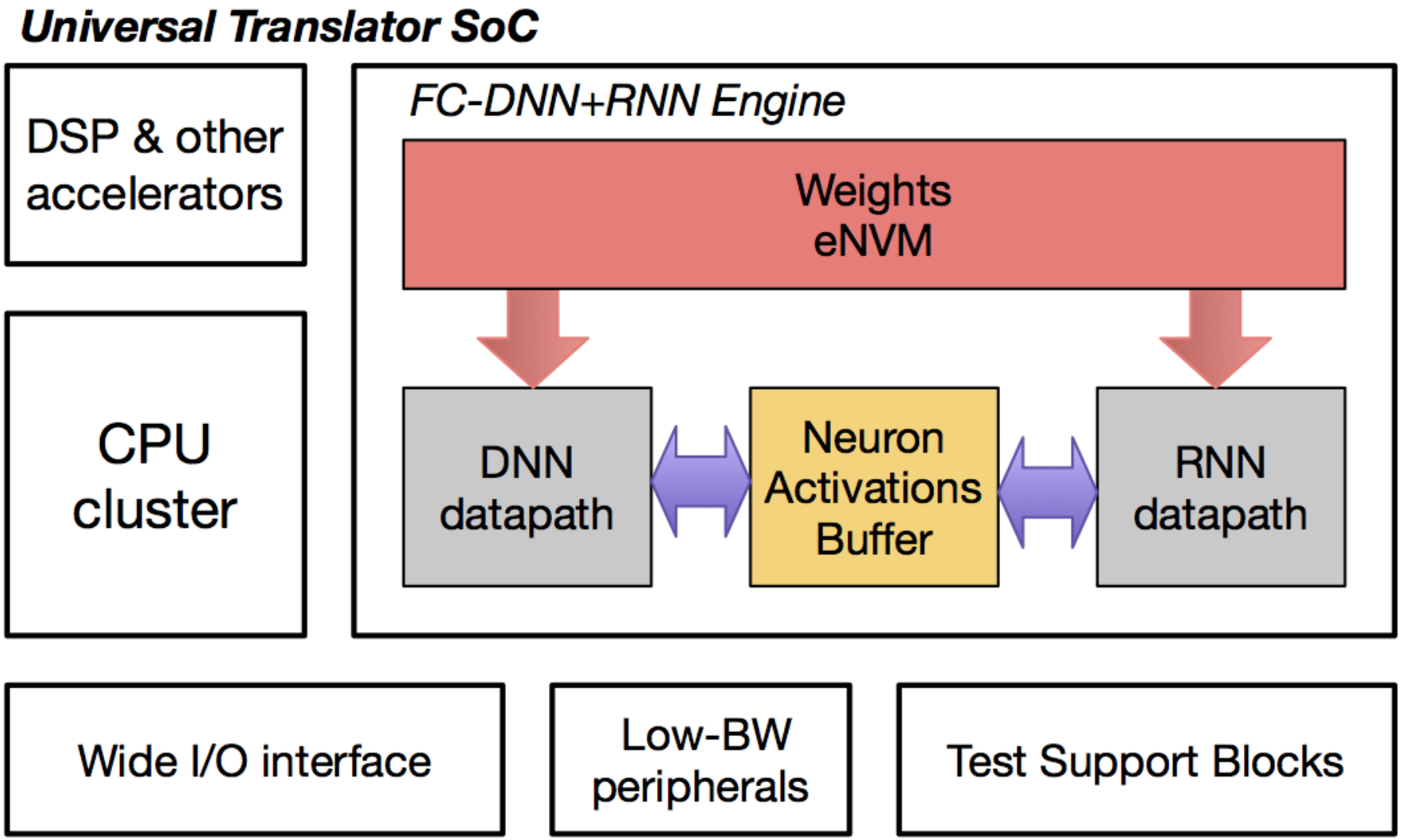
2018 and Beyond: Visualization and Debugging for NMT



2018 and Beyond: Research Models



2018 and Beyond: Hardware for NMT



Thank you.