COMP4901K/Math4824B
Machine Learning for Natural Language Processing

Lecture 15: Sequence to Sequence Learning

Instructor: Yangqiu Song

Slides credits: Richard Socher, Vagelis Hristidis

Sequence to Sequence

- Speech recognition

http://nlp.stanford.edu/courses/lsa352/

Sequence to Sequence

- Question answering

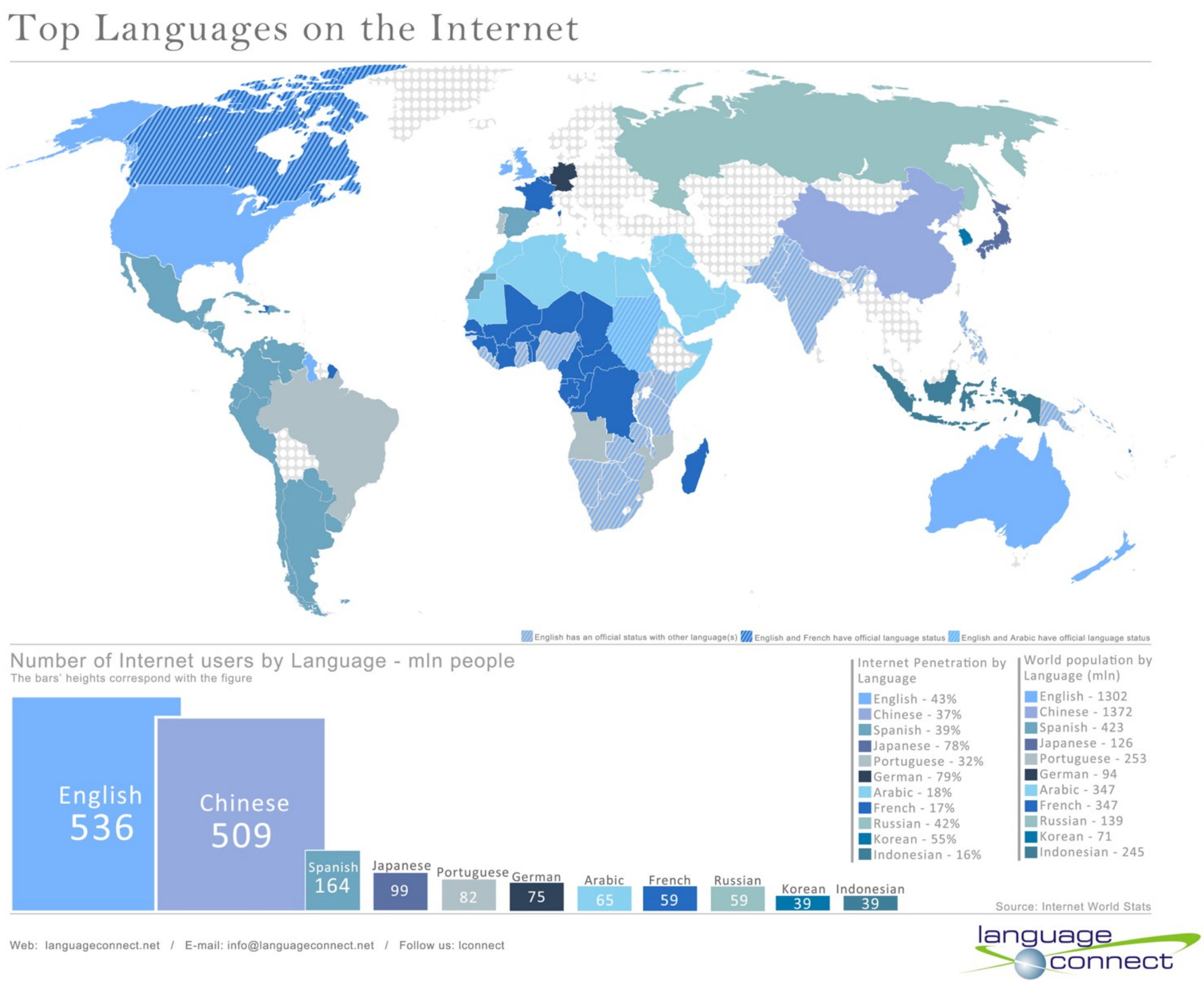
Sequence to Sequence

- Machine translation

مرحبا بكم في درس التعلم العميق

Welcome to the deep learning class

7 billion people, 7000 languages



Machine Translation

- **Machine Translation (MT)** is the task of translating a sentence *x* from one language (the source language) to a sentence *y* in another language (the target language).

- **1950s: Early Machine Translation**
  - Mostly Russian → English (motivated by the Cold War!)
  - Systems were mostly rule-based, using a bilingual dictionary to map Russian words to their English counterparts

Source: https://youtu.be/K-HfpsHPmvw

1990s-2010s: Statistical Machine Translation

- Core idea: Learn a probabilistic model from data
- Suppose we're translating Language X → English.
- We want to find best English sentence *y*, given Language X sentence *x*

$$\operatorname{argmax}_y P(y|x)$$

- Use Bayes Rule to break this down into two components to be learnt separately:

$$= \operatorname{argmax}_y P(x|y) P(y)$$

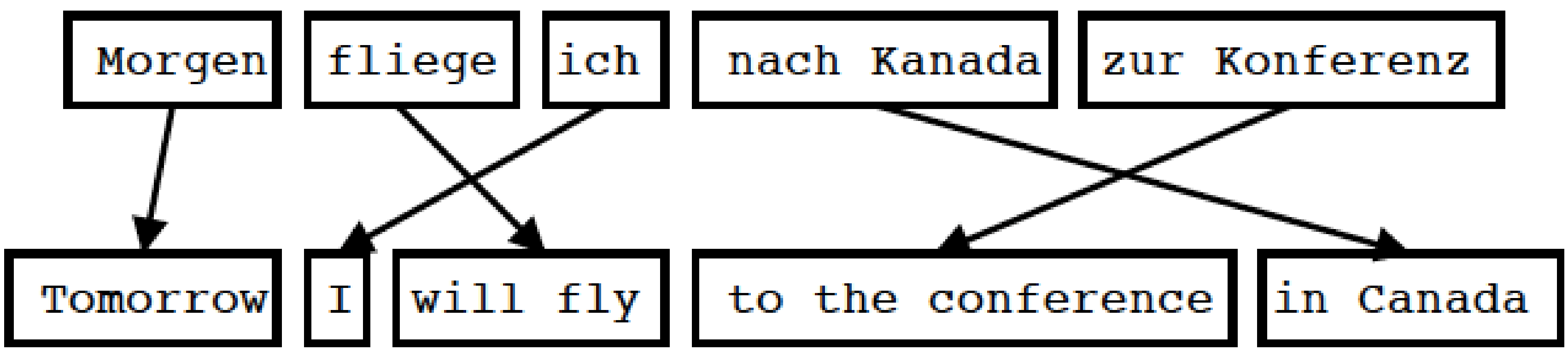
Translation Model
Models how words and phrases should be translated.
Learnt from parallel data.

Language Model
Models how to write good English.
Learnt from monolingual data.

Statistical Machine Translation

- Translation model

- Input is Segmented in Phrases
- Each Phrase is Translated into English
- Phrases are Reordered



Koehn 2004

Statistical Machine Translation

- Language Model

Goal of the Language Model: Detect good English

For Example: Trigram Model

```
Mary did not slap the green witch

Mary  =>  p(Mary)
Mary did   =>  p(did|Mary)
Mary did not  =>  p(not|Mary did)
     did not slap  =>  p(slap|did not)
         not slap the  =>  p(the|not slap)
             slap the green  =>  p(green|slap the)
                  the green witch  =>  p(witch|the green)
```

Knight and Koehn 2003

1990s-2010s: Statistical Machine Translation

$$\text{argmax}_y P(x|y)P(y)$$

Question:
How to compute this argmax?

Translation Model

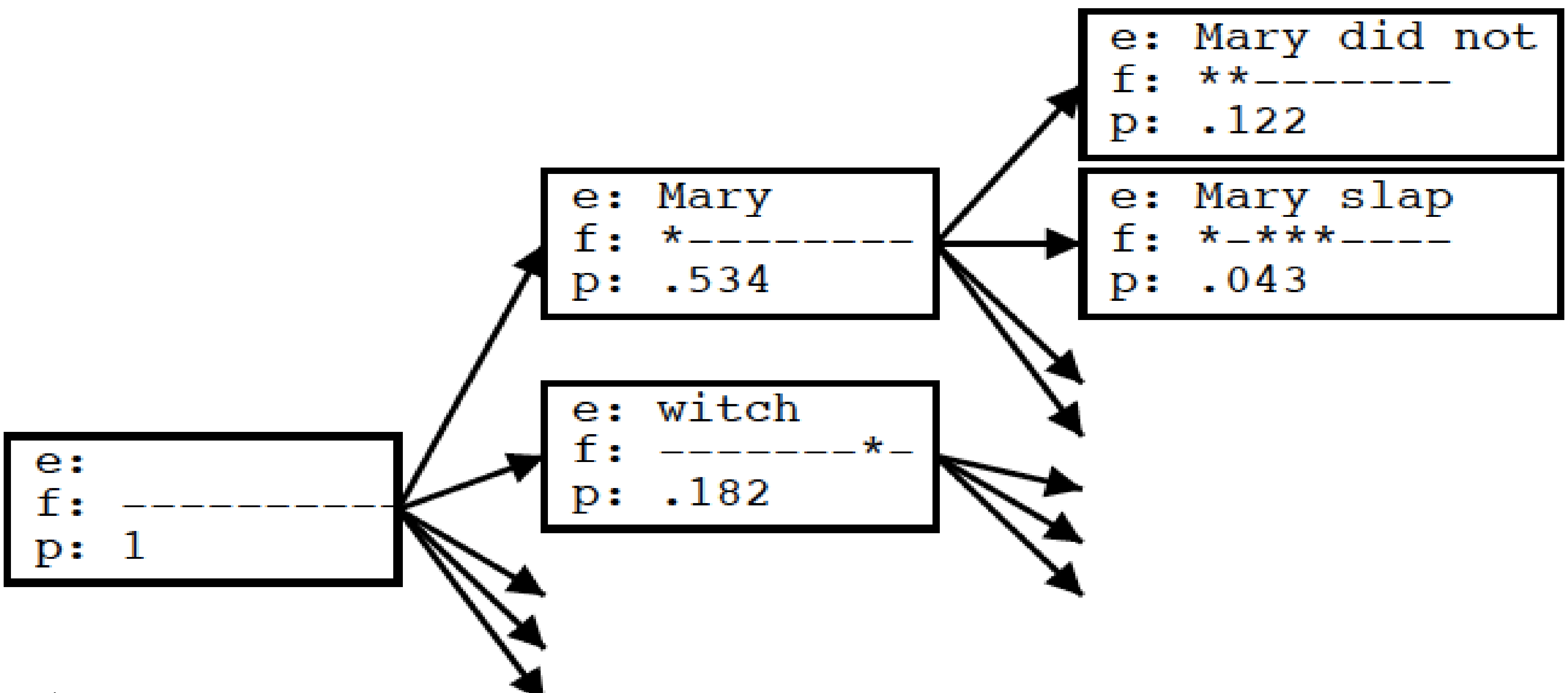
Language Model

- We could enumerate every possible *y* and calculate the probability? → Too expensive!
- **Answer:** Use a heuristic search algorithm to gradually build up the translation, discarding hypotheses that are too low probability

Statistical Machine Translation

- Decoding

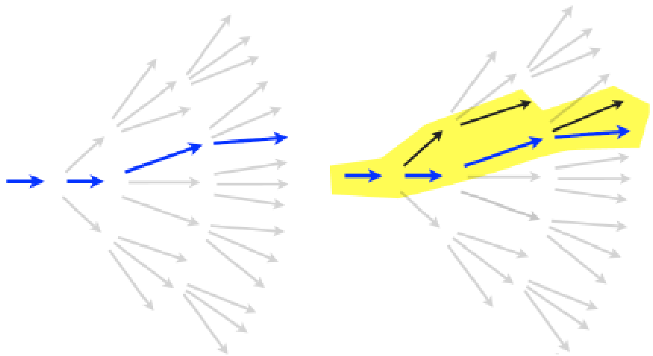
Goal of the decoding algorithm: Put models to work, perform the actual translation



Koehn 2004

Decoding

- Global solution: Dynamic programming
- Approximate solutions: Beam inference
  - At each position keep the top k complete sequences
  - Extend each sequence in each local way
  - The extensions compete for the k slots at the next position



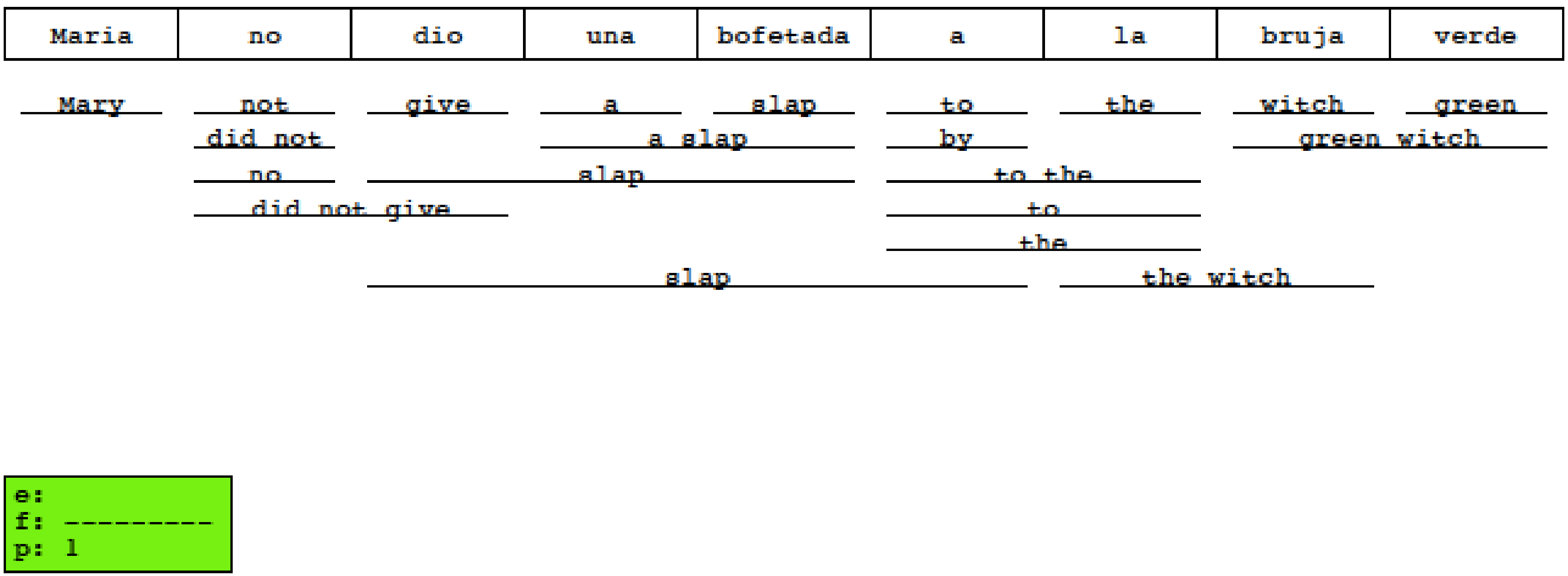
(a) Greedy (b) Beam Search

- Advantages
  - Fast; beam sizes of 3-5 are almost as good as exact inference in many cases
  - Easy to implement (no dynamic programming required)
- Disadvantage
  - Inexact: the globally best sequence can fall off the beam

Statistical Machine Translation

- Decoding

Goal of the decoding algorithm: Put models to work, perform the actual translation



Koehn 2004

Statistical Machine Translation

- Decoding

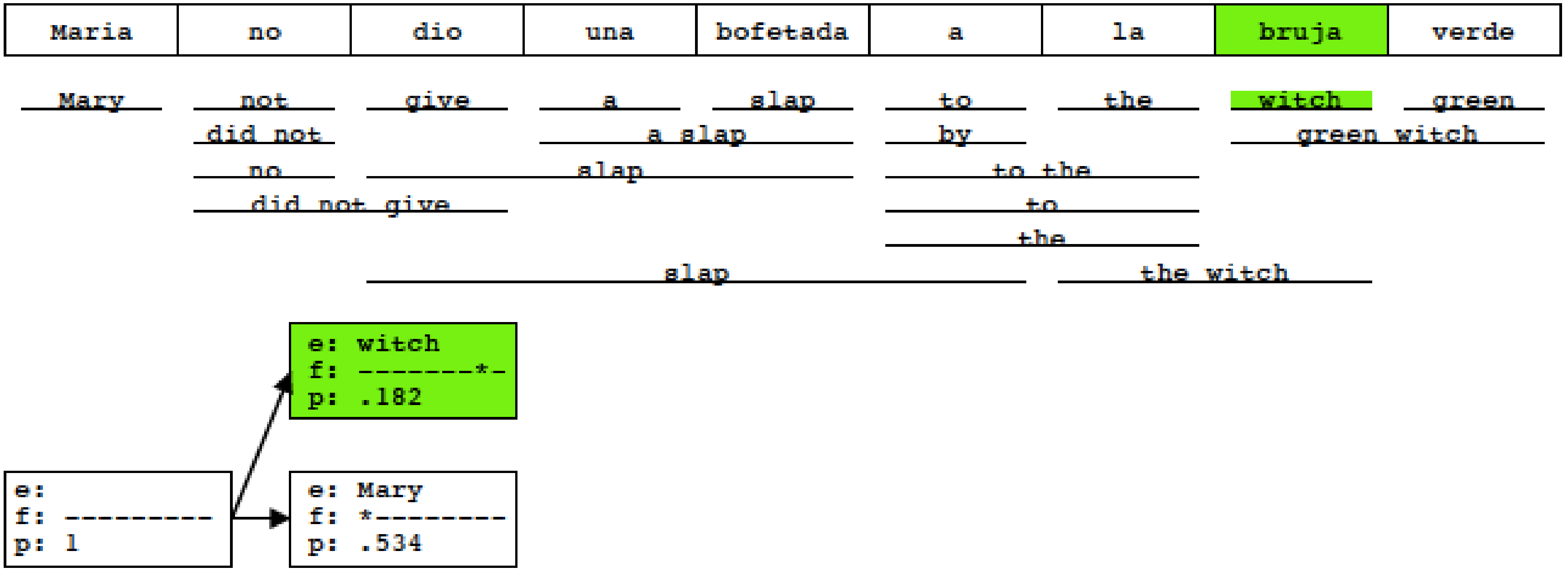
Goal of the decoding algorithm: Put models to work, perform the actual translation

Koehn 2004

Statistical Machine Translation

- Decoding

Goal of the decoding algorithm: Put models to work, perform the actual translation

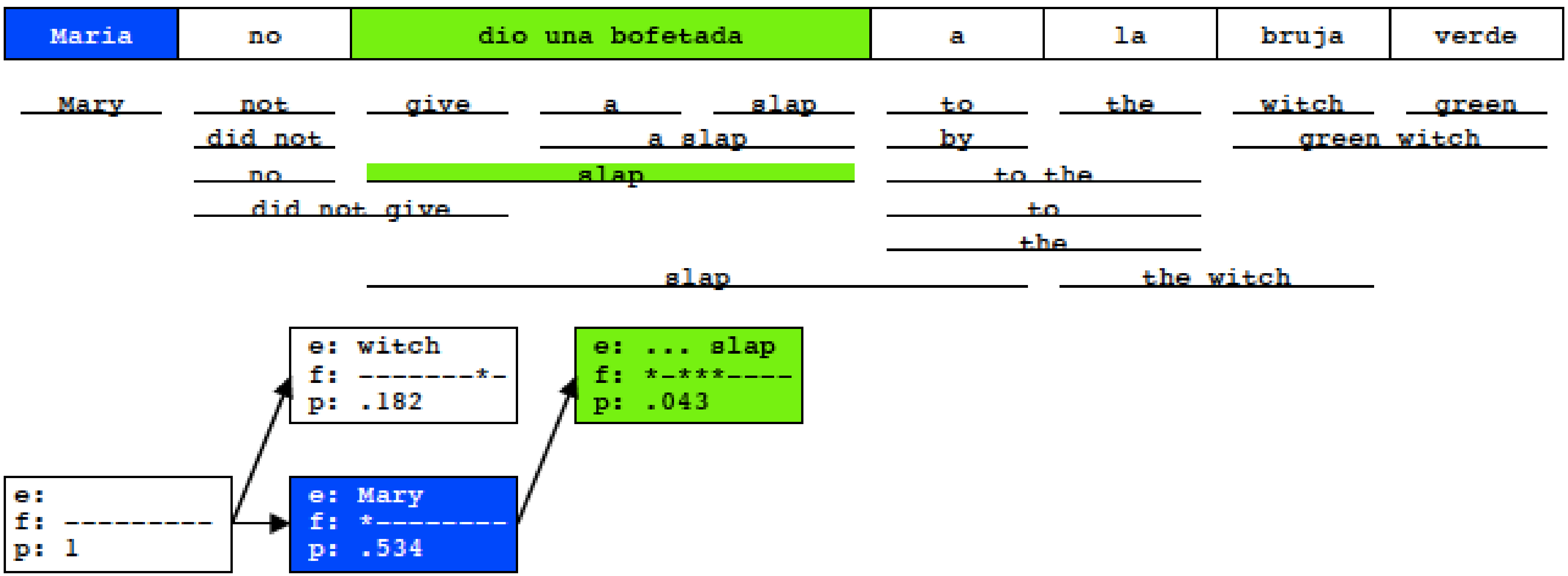


Koehn 2004

Statistical Machine Translation

- Decoding

Goal of the decoding algorithm: Put models to work, perform the actual translation

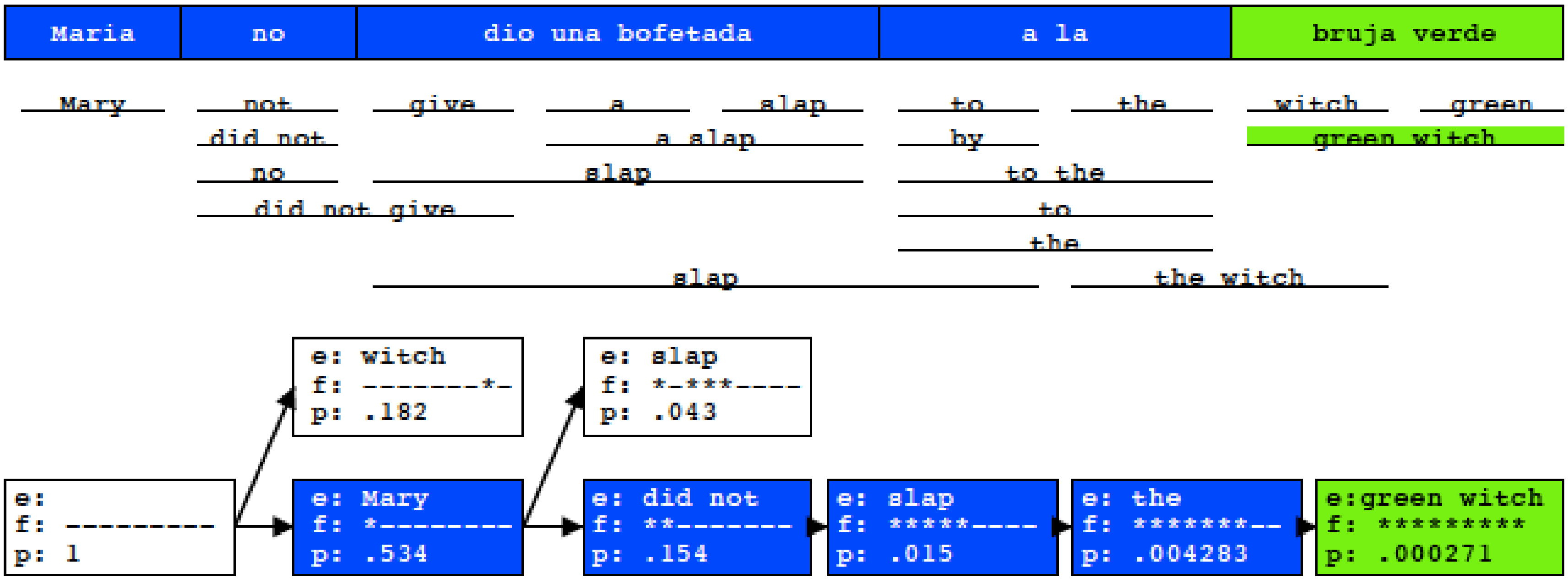


Koehn 2004

Statistical Machine Translation

- Decoding

Goal of the decoding algorithm: Put models to work, perform the actual translation



Koehn 2004

1990s-2010s: Statistical Machine Translation

- SMT is a huge research field
- The best systems are extremely complex
  - Hundreds of important details we haven't mentioned here
  - Systems have many separately-designed subcomponents
  - Lots of feature engineering
    - Need to design features to capture particular language phenomena
  - Require compiling and maintaining extra resources
    - Like tables of equivalent phrases
  - Lots of human effort to maintain
    - Repeated effort for each language pair!

Neural Machine Translation

- Neural Machine Translation (NMT) is a way to do Machine Translation with a *single neural network*

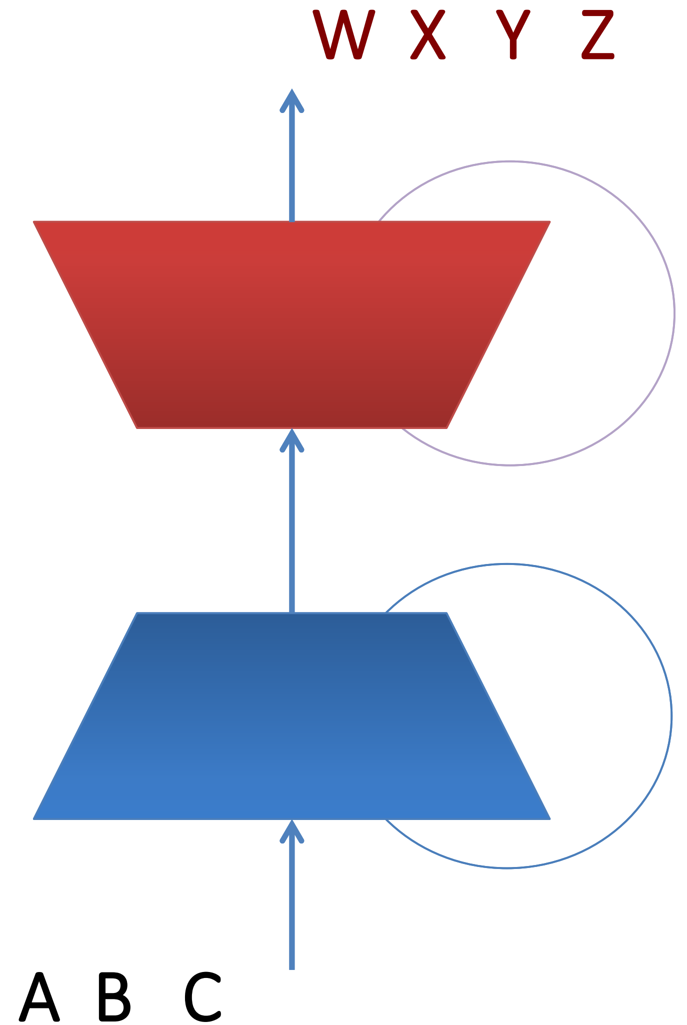
- The neural network architecture is called sequence-to-sequence (aka seq2seq) and it involves *two* RNNs.

Sutskever et al.,2014

Sequence to Sequence Learning with Neural Networks

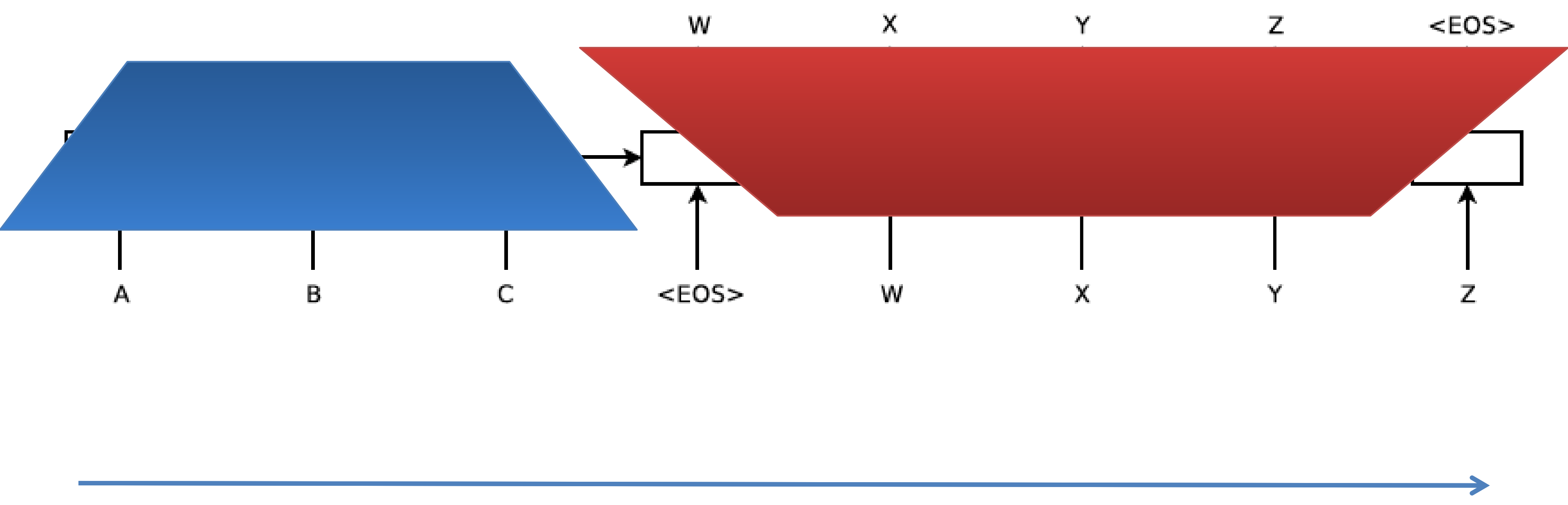
Neural Machine Translation

- Model



Neural Machine Translation

- Model



Sutskever et al. 2014

Neural Machine Translation (NMT)

- The sequence-to-sequence model is an example of a **Conditional Language Model**.
  - **Language Model** because the decoder is predicting the next word of the target sentence *y*
  - **Conditional** because its predictions are *also* conditioned on the source sentence *x*

- NMT directly calculates $P(y|x)$

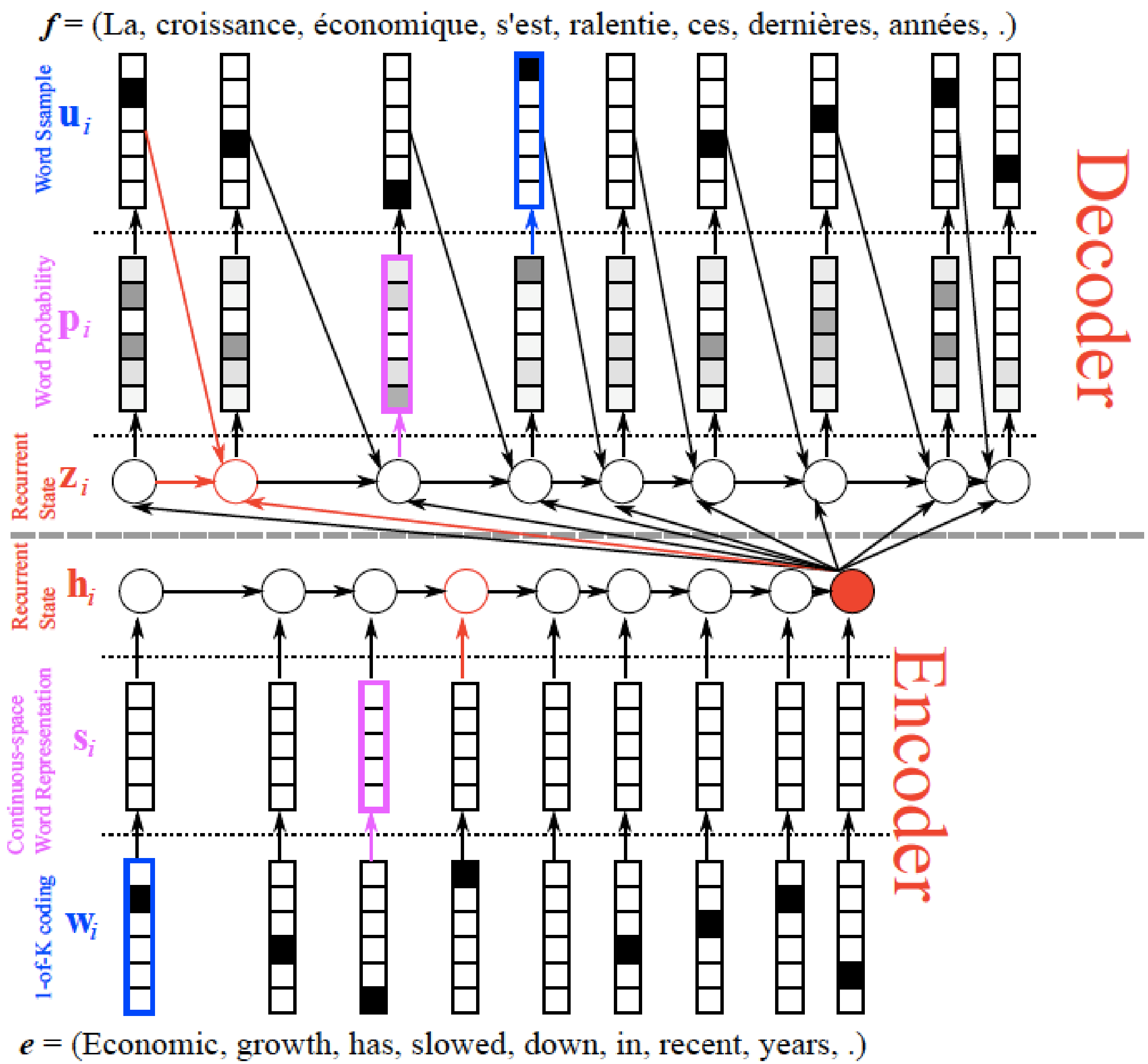
$$P(y|x) = P(y_1|x)\, P(y_2|y_1, x)\, P(y_3|y_1, y_2, x) \ldots, \underline{P(y_T|y_1, \ldots, y_{T-1}, x)}$$

Probability of next target word, given target words so far and source sentence *x*

- **Question**: How to train a NMT system?
- **Answer**: Get a big parallel corpus…

Neural Machine Translation

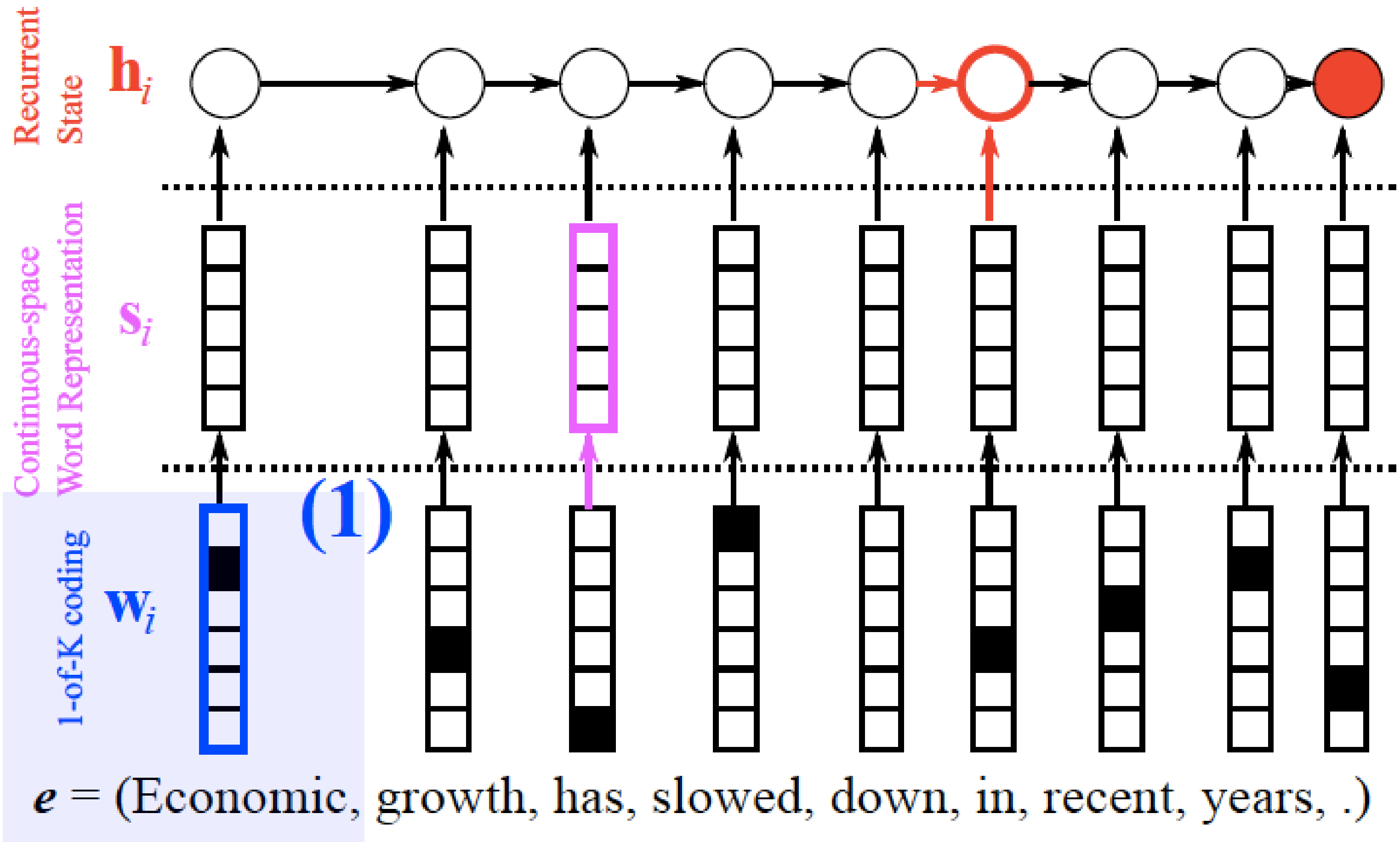
- Model



Cho: From Sequence Modeling to Translation

Neural Machine Translation

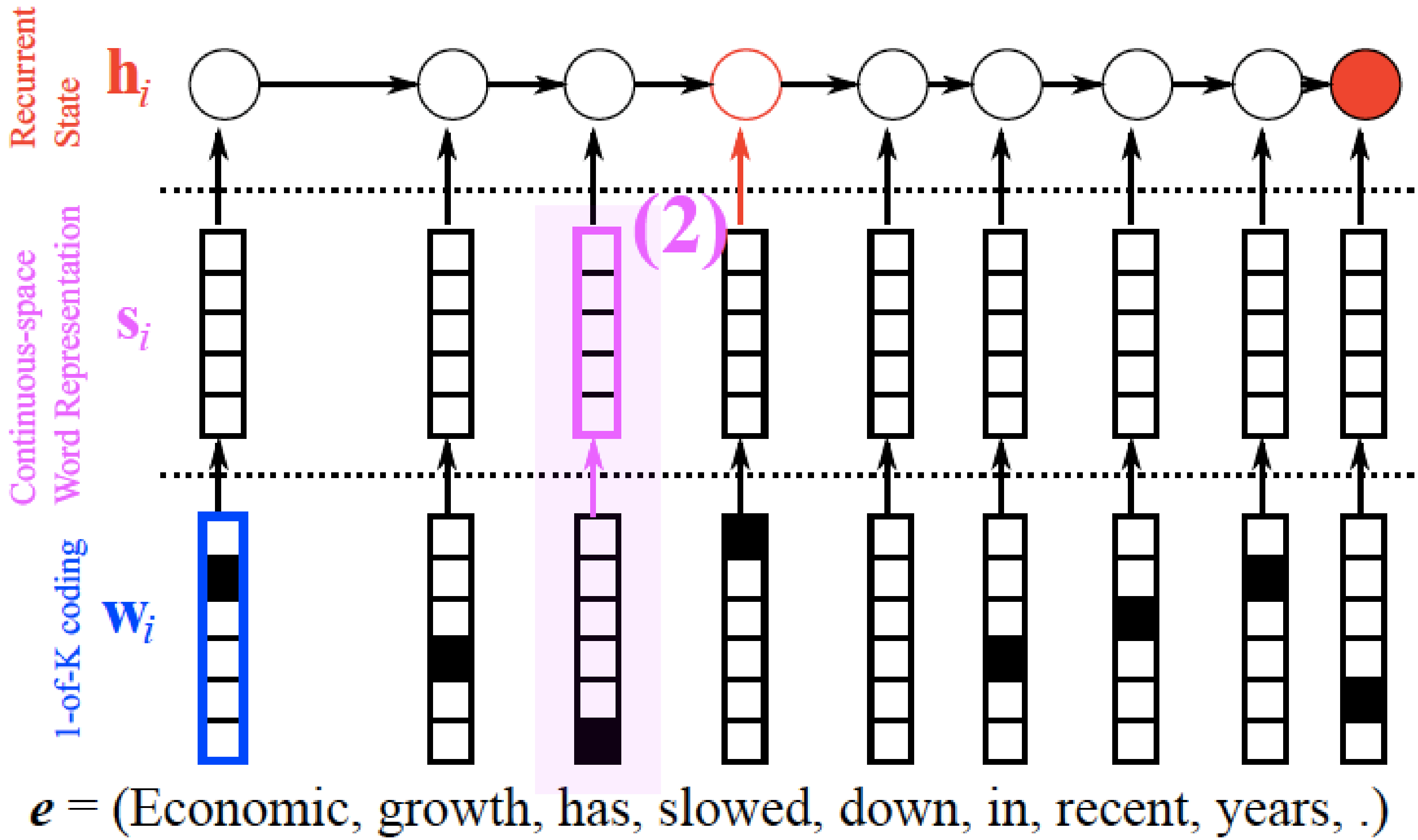
- Model- *encoder*



Cho: From Sequence Modeling to Translation

Neural Machine Translation

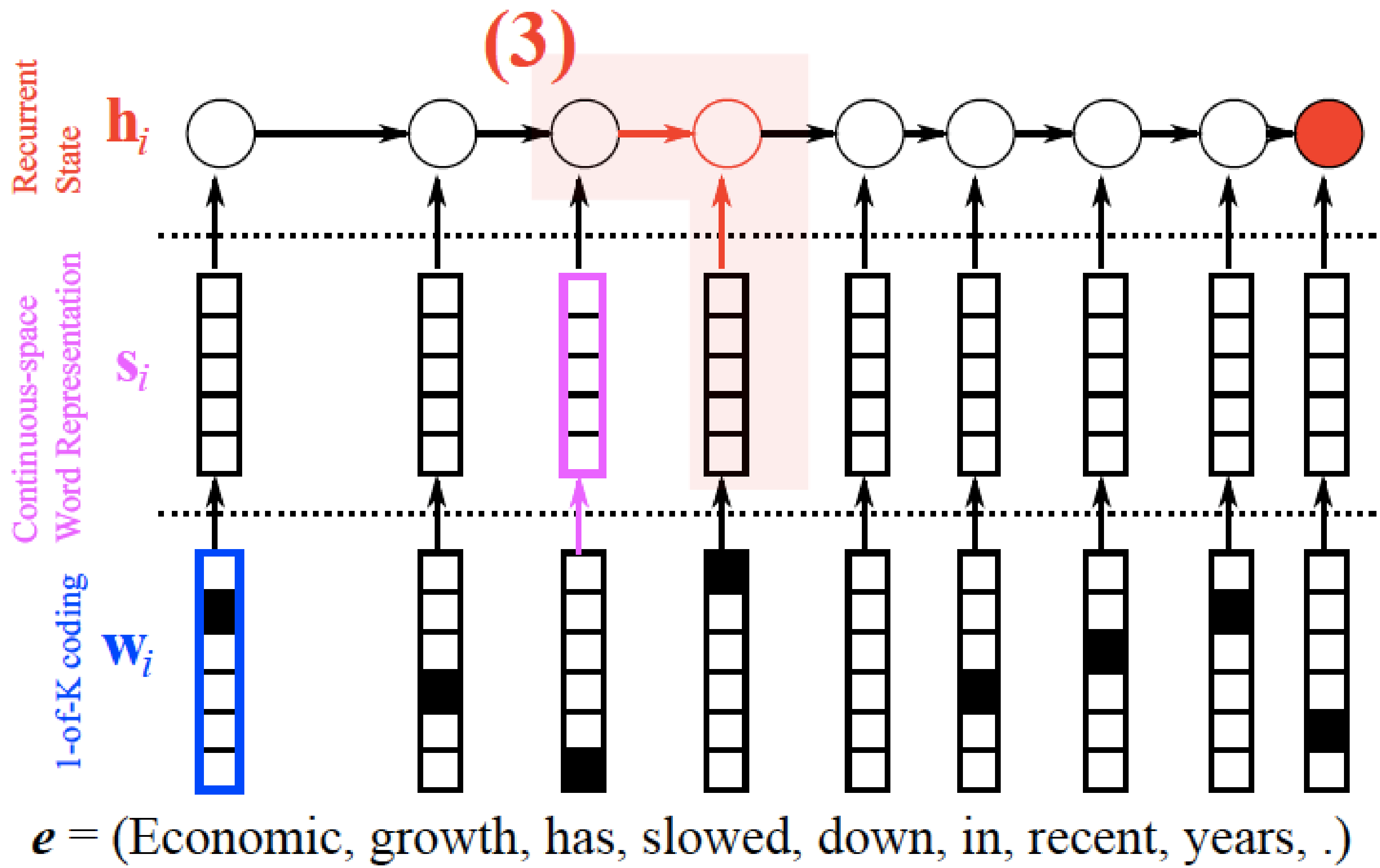
- Model- *encoder*



Cho: From Sequence Modeling to Translation

Neural Machine Translation

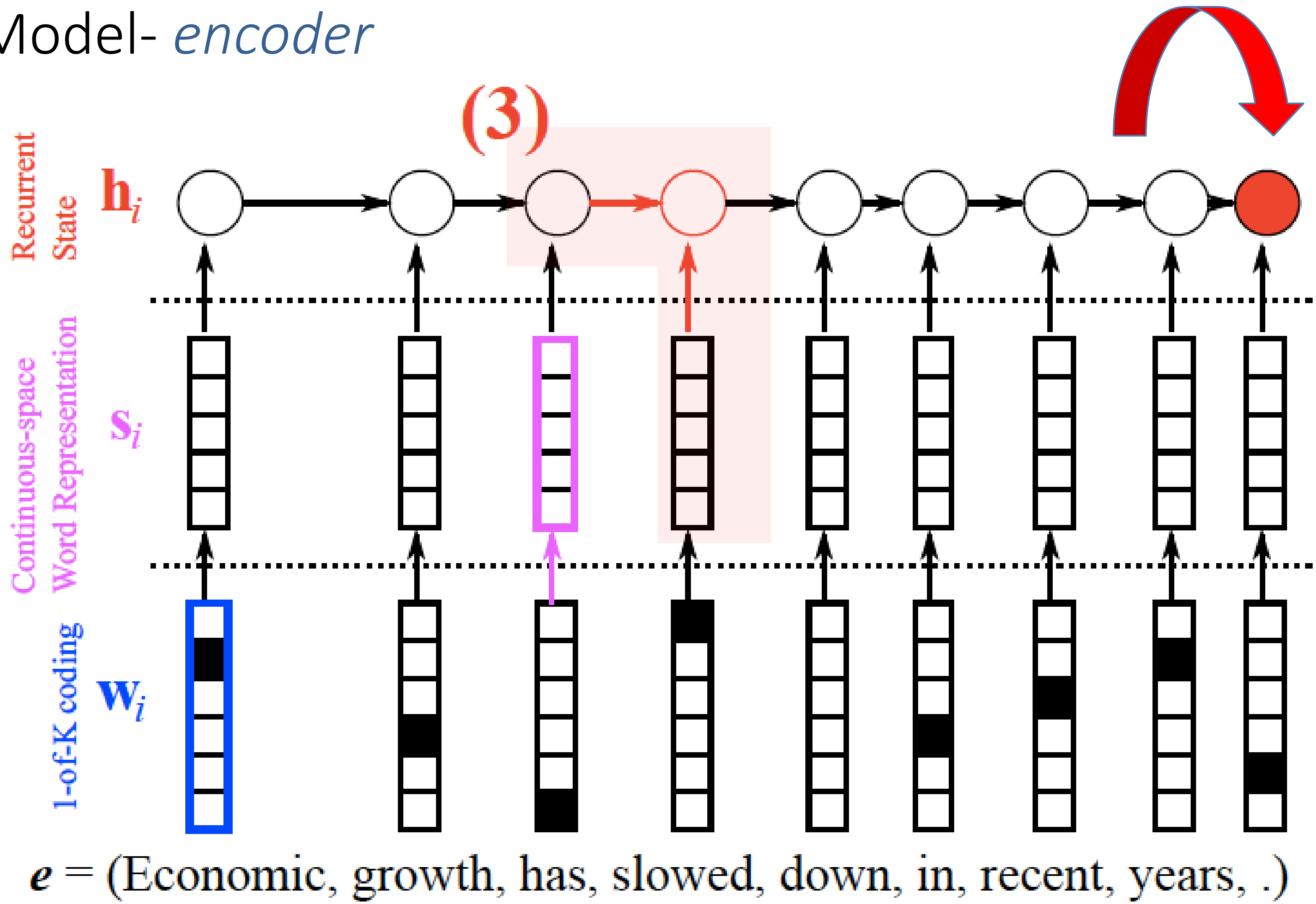
- Model- *encoder*



Cho: From Sequence Modeling to Translation

Neural Machine Translation

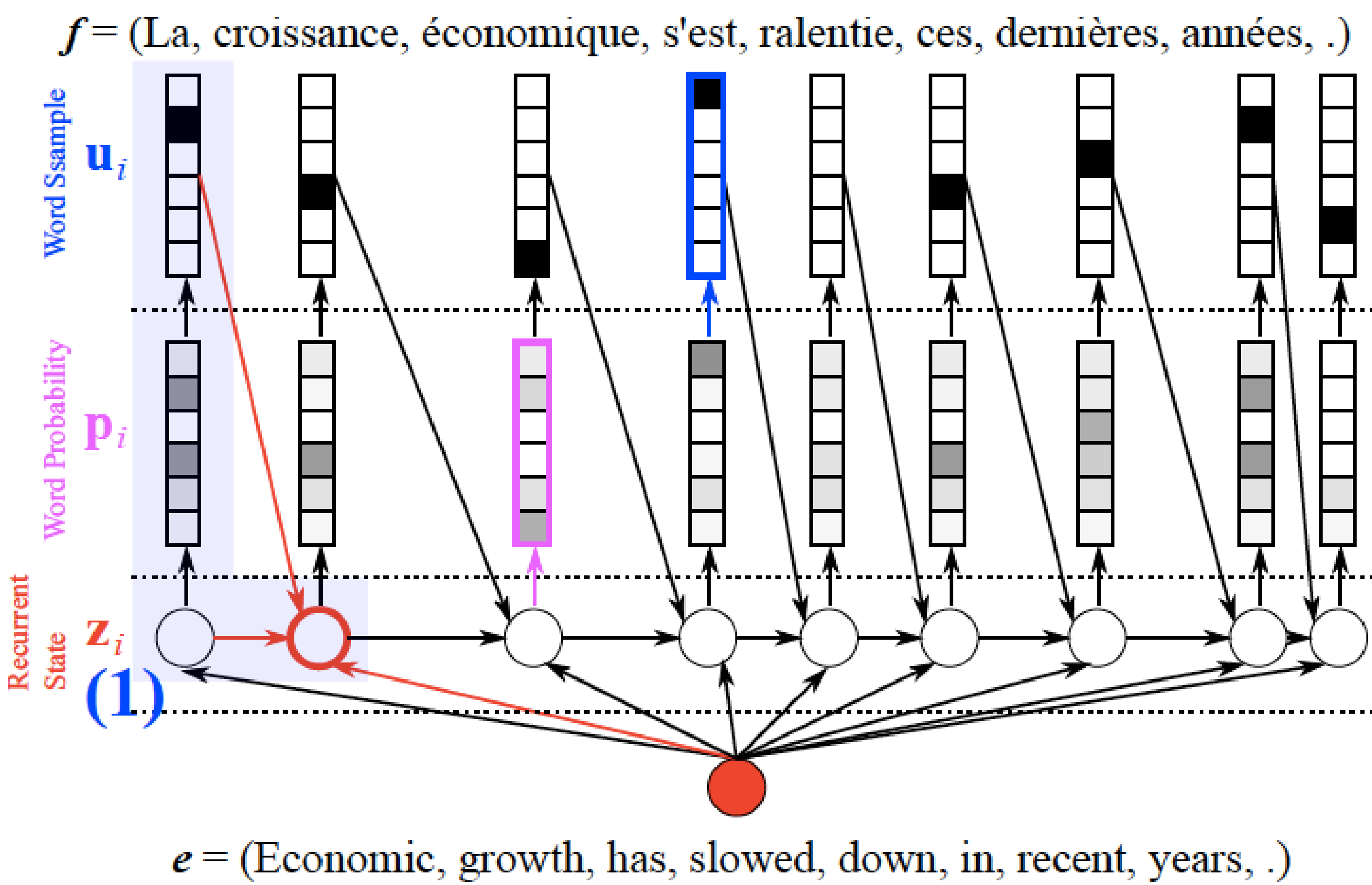
- Model- *encoder*



Cho: From Sequence Modeling to Translation

Neural Machine Translation

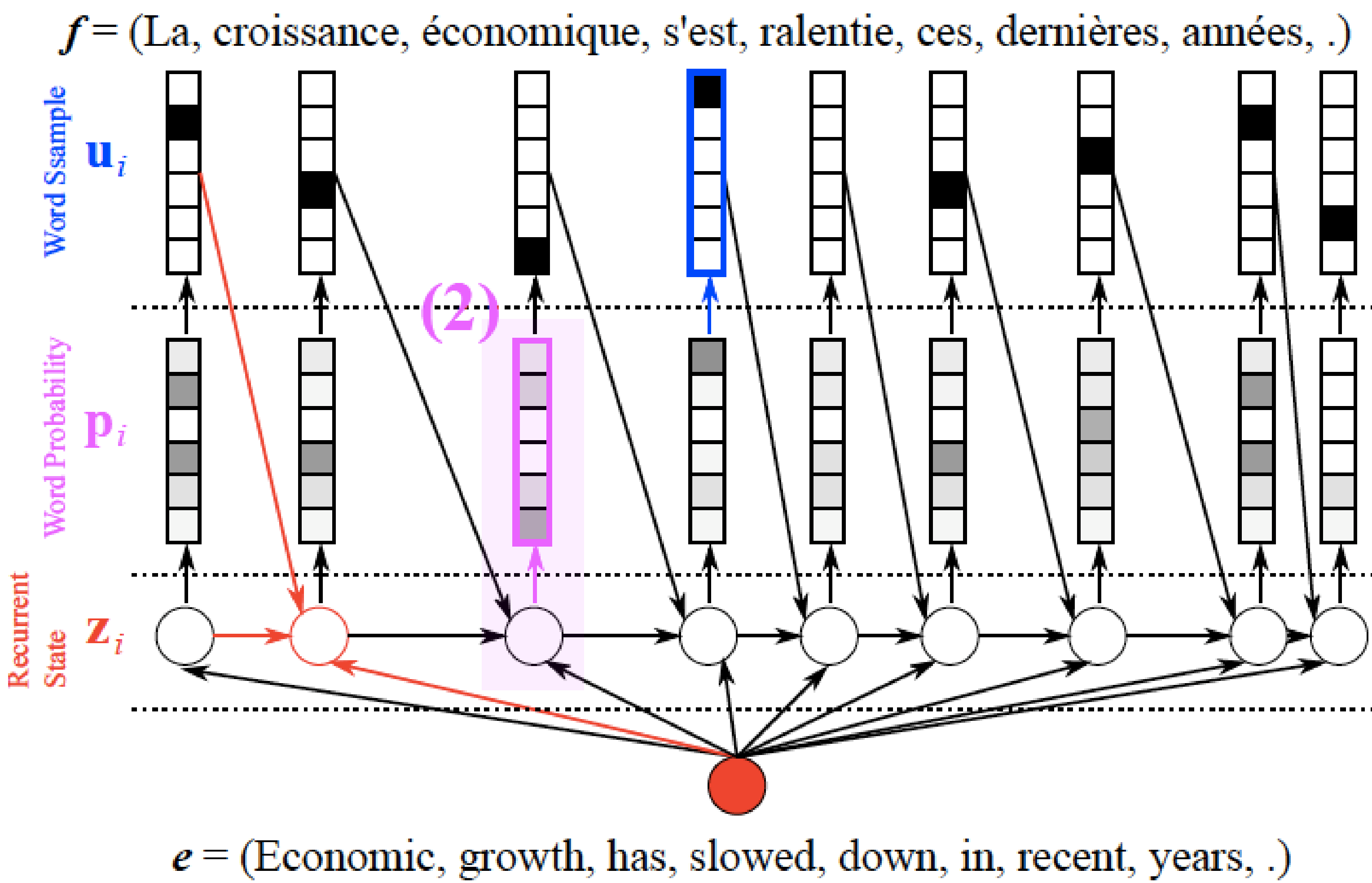
- Model- *decoder*



Cho: From Sequence Modeling to Translation

Neural Machine Translation

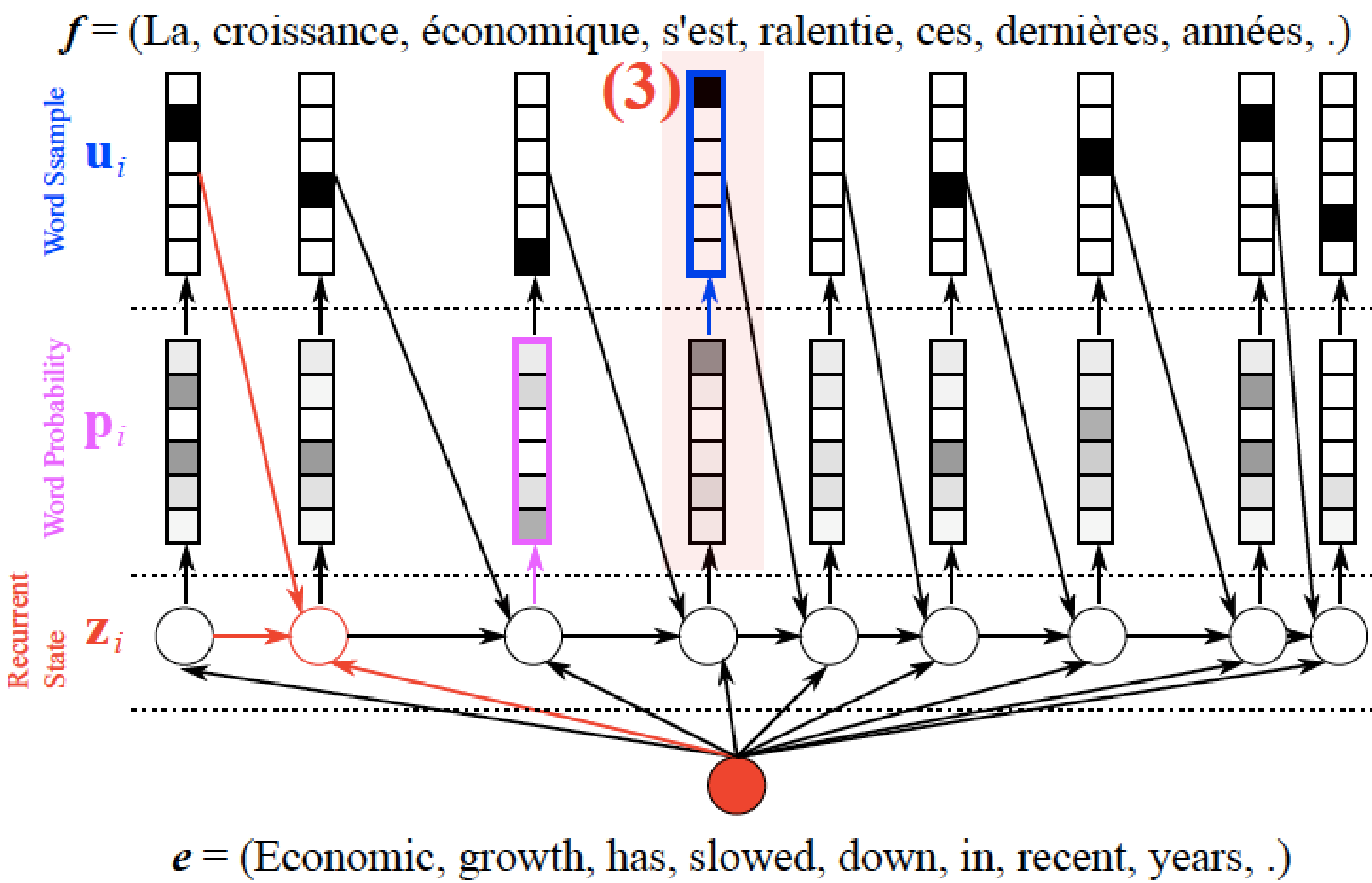
- Model- *decoder*



Cho: From Sequence Modeling to Translation

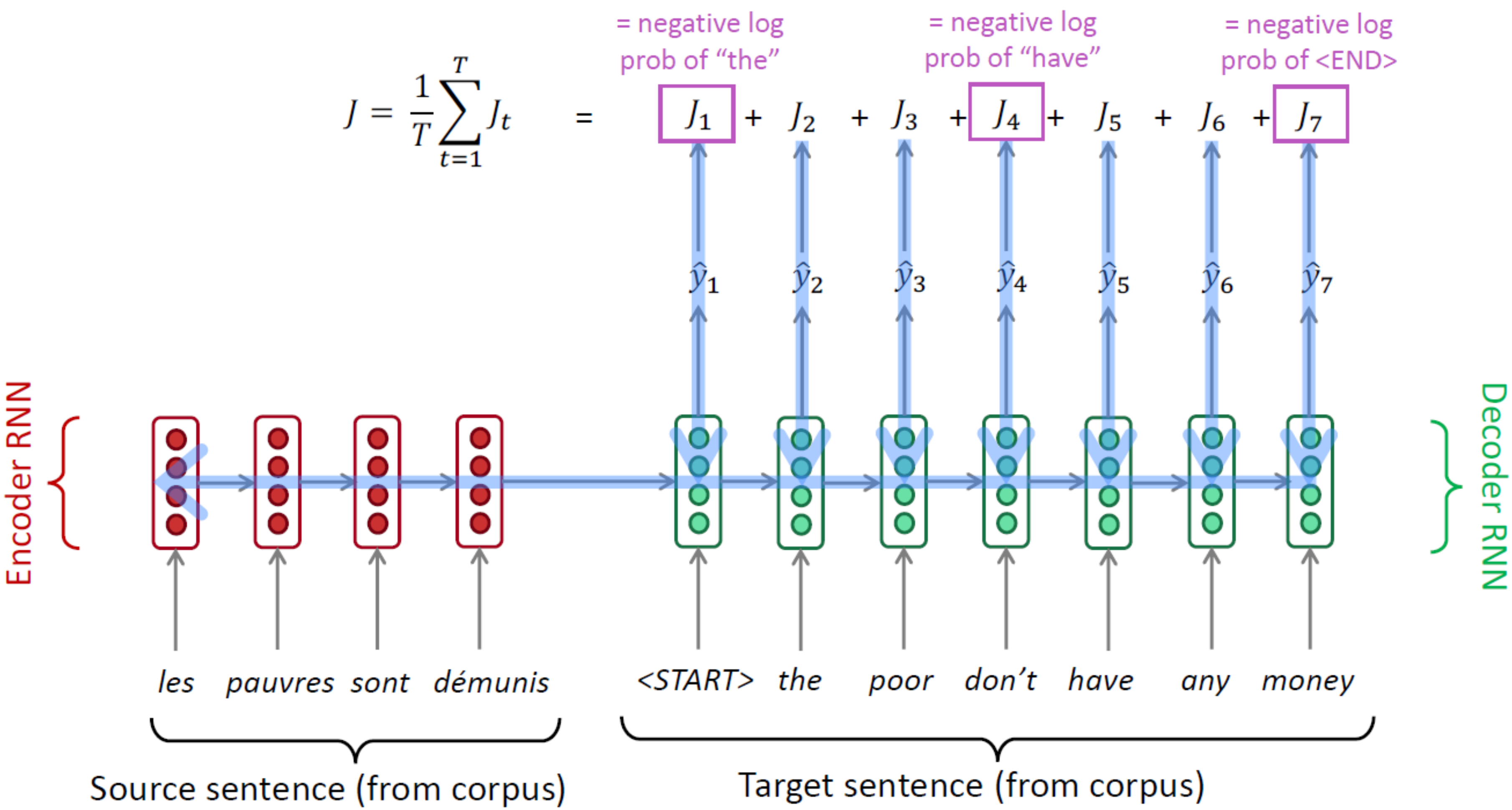
Neural Machine Translation

- Model- *decoder*



Cho: From Sequence Modeling to Translation

Training a Neural Machine Translation system



Seq2seq is optimized as a **single system.**
Backpropagation operates *"end to end"*.

Advantages of NMT

Compared to SMT, NMT has many advantages:

- Better performance
  - More fluent
  - Better use of context
  - Better use of phrase similarities
- A single neural network to be optimized end-to-end
  - No subcomponents to be individually optimized
- Requires much less human engineering effort
  - No feature engineering
  - Same method for all language pairs

Disadvantages of NMT?

Compared to SMT:

- NMT is less interpretable
  - Hard to debug
- NMT is difficult to control
  - For example, can’t easily specify rules or guidelines for translation
  - Safety concerns!

Evaluation (Machine Translation)

- BLEU (bilingual evaluation understudy) (Papineni et al. (2002))
  - BLEU compares the machine-written translation to one or several human-written translation(s), and computes a similarity score based on:
    - *n*-gram precision (usually up to 3 or 4-grams)
    - Penalty for too-short system translations
  - BLEUs output is always a number between 0 and 1
    - 1 means identical to the reference translations

BLEU is useful but imperfect

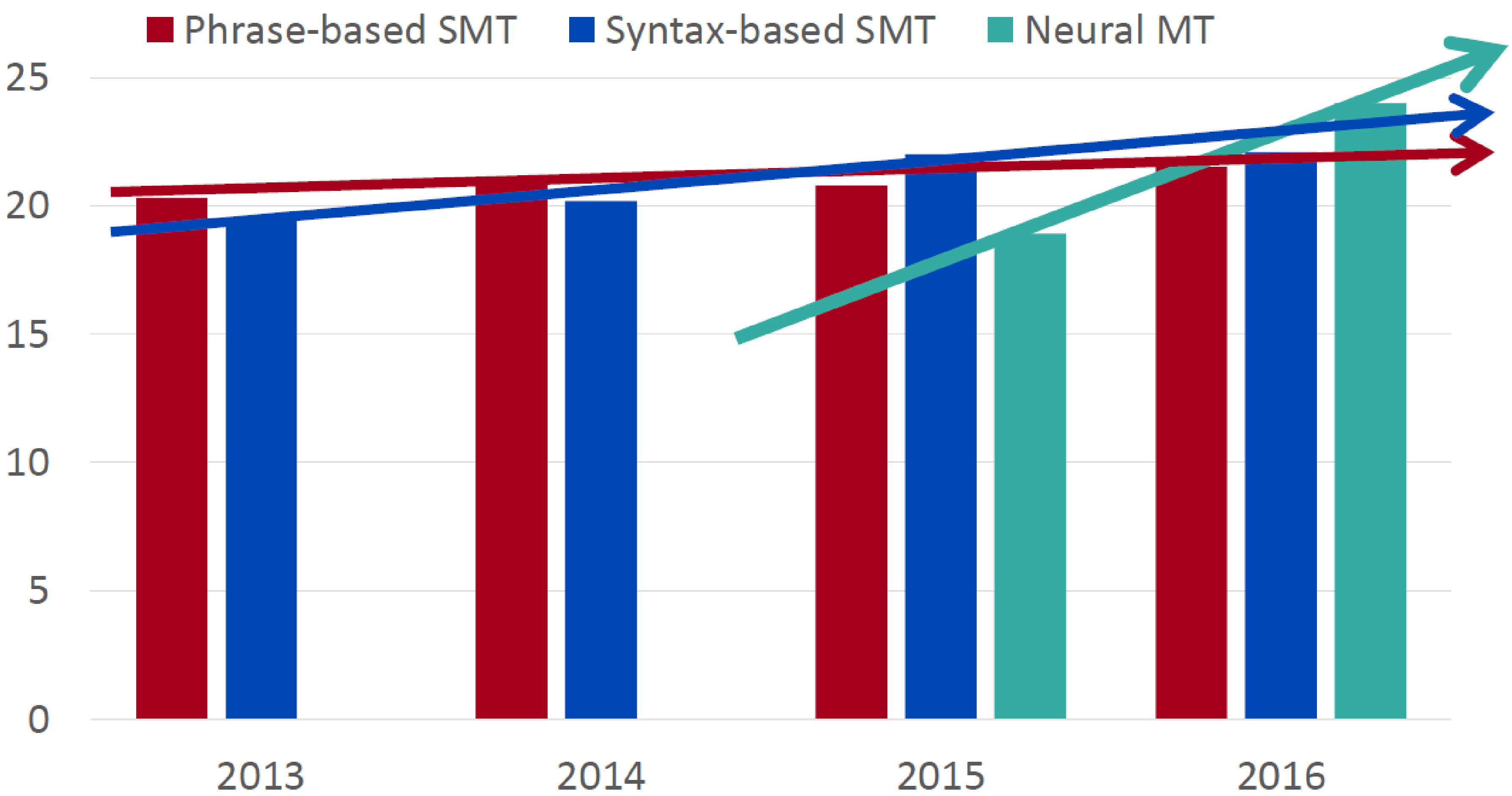
- BLEU was one of the first metrics to claim a high correlation with human judgements of quality, and remains one of the most popular automated and inexpensive metrics

- There are many valid ways to translate a sentence

- So a good translation can get a poor BLEU score because it has low *n*-gram overlap with the human translation L

MT progress over time

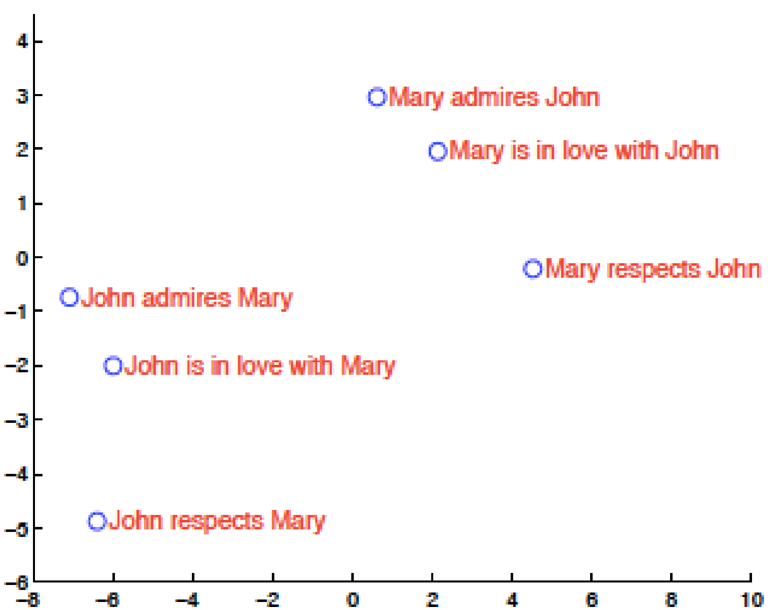
[Edinburgh En-De WMT newstest2013 Cased BLEU; NMT 2015 from U. Montréal]

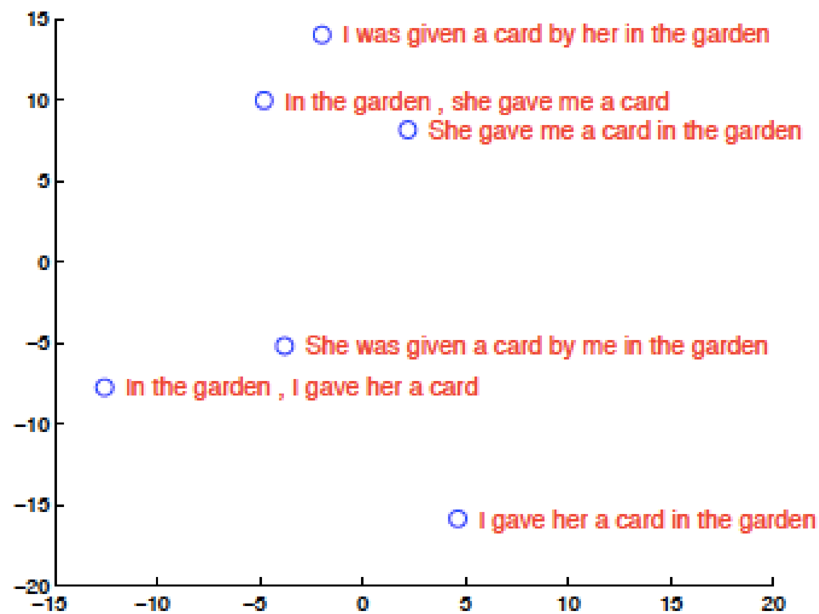


Source: http://www.meta-net.eu/events/meta-forum-2016/slides/09_sennrich.pdf

NMT Analysis

- Model Analysis

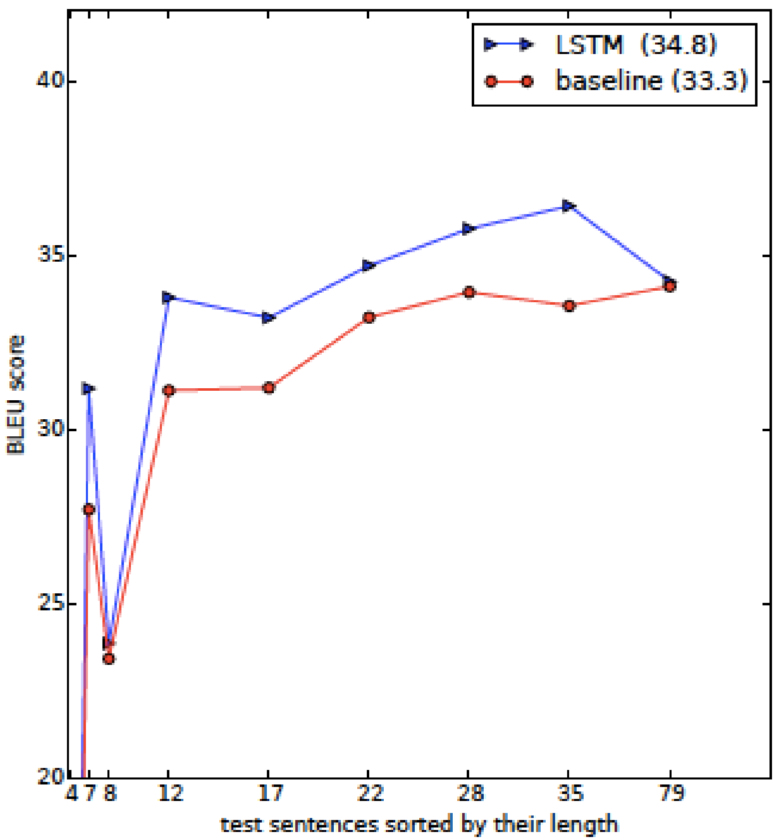


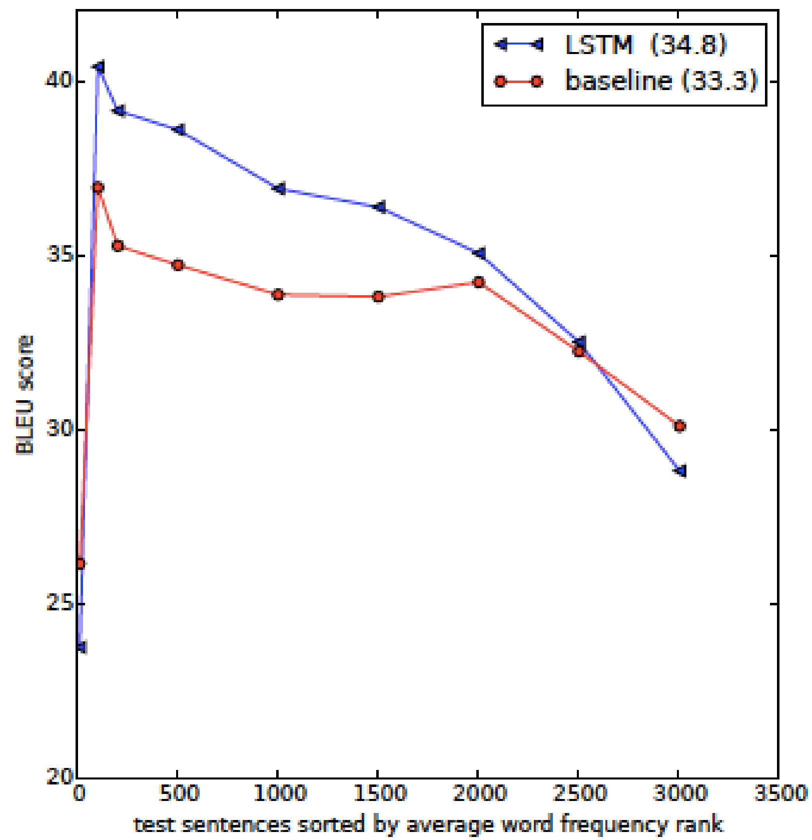


Sutskever et al. 2014

NMT Analysis

- Long sentences





Sutskever et al. 2014

NMT: the biggest success story of NLP Deep Learning

- Neural Machine Translation went from a fringe research activity in **2014** to the leading standard method in **2016**
  - **2014**: First seq2seq paper published
  - **2016**: Google Translate switches from SMT to NMT

- This is amazing!
  - **SMT** systems, built by hundreds of engineers over many years, outperformed by NMT systems trained by a handful of engineers in a few months

So is Machine Translation solved?

- **Nope!**
- Many difficulties remain:
  - Out-of-vocabulary words
  - Domain mismatch between train and test data
  - Maintaining context over longer text
  - Low-resource language pairs

- (Old) Bad Examples
  - http://languagelog.ldc.upenn.edu/nll/?p=35120#more-35120
  - https://hackernoon.com/bias-sexist-or-this-is-the-way-it-should-be-ce1f7c8c683c

NMT research continues

- NMT is the **flagship task** for NLP Deep Learning
  - NMT research has pioneered many of the recent innovations of NLP Deep Learning

- In **2018**: NMT research continues to thrive
  - Researchers have found *many, many* improvements to the "vanilla" seq2seq NMT system we've presented today