

بسم الله الرحمن الرحيم

# پردازش زبانهای طبیعی

جلسه ۲۳

احسان الدین عسگری

خرداد ۱۴۰۳

<http://language.ml/>

دانشکده مهندسی کامپیوتر

آزمایشگاه پردازش هوشمند متن و زبان و علوم انسانی محاسباتی

asgari@berkeley.edu



# ماژولهای درس



مسائل کاربردی دیگر



تولید متن



طبقه‌بندی کلمه  
و پرسش و پاسخ



طبقه‌بندی متن



مدلهای زبانی



آشنایی با متن و کلمات  
روشهای پیش‌پردازش

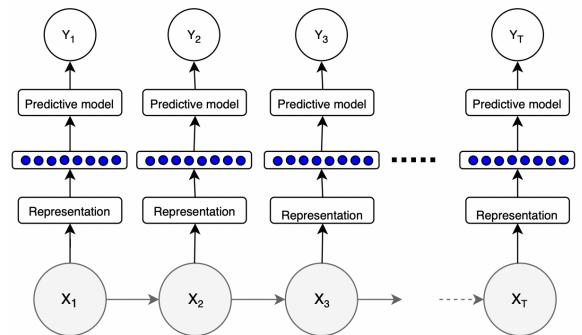


ترجمه ماشيني

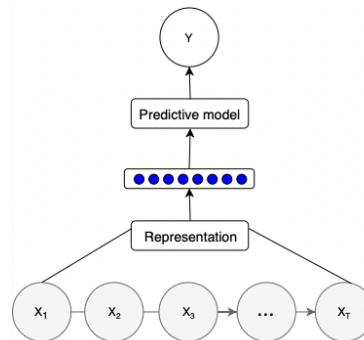
# Machine Translation



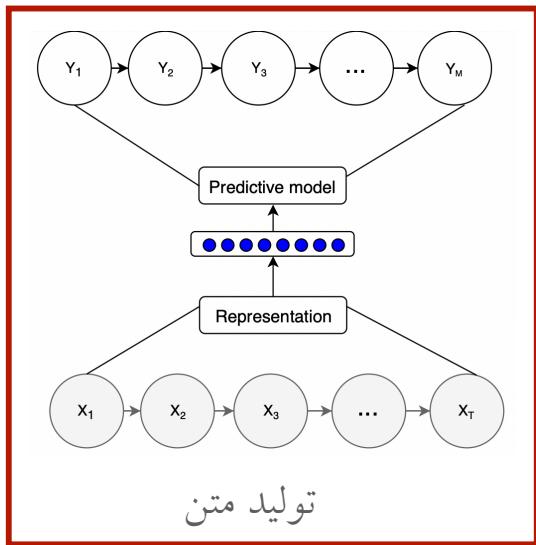
# مسائل پردازش زبان



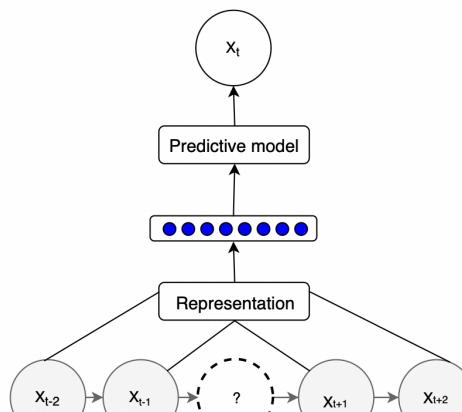
برچسب زنی



طبقه‌بندی



تولید متن

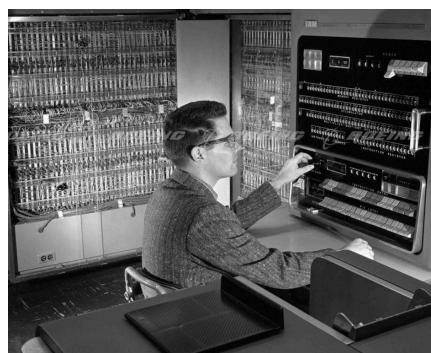


مدل زبانی

- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

# تاریخچه ترجمه ماشینی

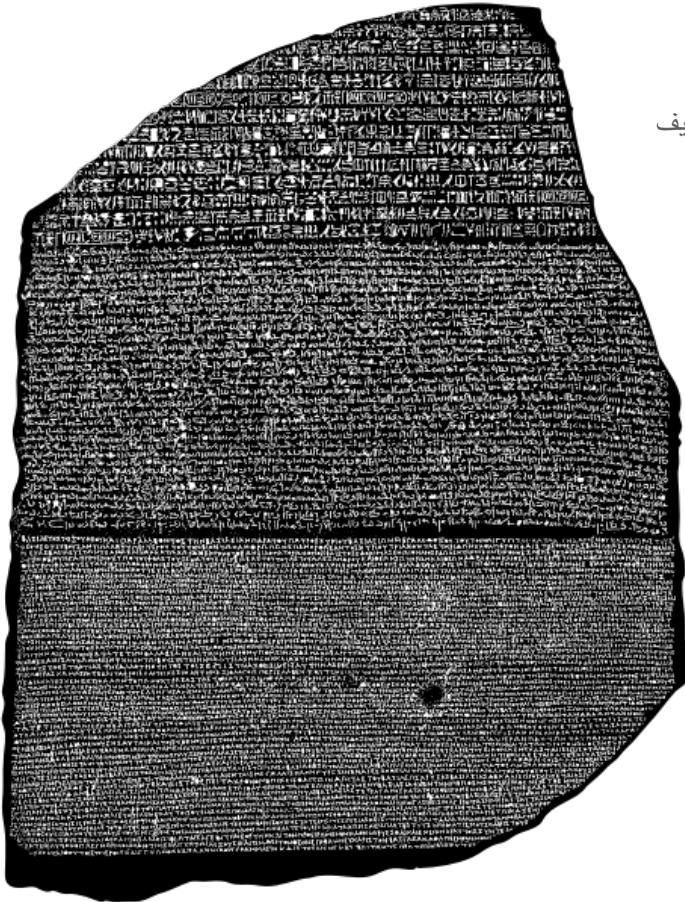
ابویوسف الکندی (۱۸۵-۲۶۵ هـق): تکنیکهایی برای ترجمه سیستماتیک با استفاده از آمار.



۱۹۵۴ - ترجمه جملات روسی به انگلیسی در جنگ سرد.

- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

# متنهای موازی



هیروگلیف

دموتیک

یونانی باستان

سنگ رزتا - ۱۹۶ پیش از میلاد مسیح ع

- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

# هدف ترجمه ماشینی

ترجمه جمله **X** از زبان مبداء (زبان **source**) به جمله **y** از زبان مقصد (زبان **target**).

- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

X: فَرَجَعَنَاكَ إِلَى أُمّكَ كَيْ تَقَرَّ عَيْنُهَا وَلَا تَحْزَنَ (طه ٤٠)

Y: پس تو را به سوی مادرت بازگردانیدیم تا دیدهاش روشن شود و غم نخورد.

# **Review of Commonly Used Metrics in NLP**



# Metrics

## Precision, Recall, Acc, F1

- **Precision:**

*Definition:* Measures the proportion of positive identifications that were actually correct.

*Formula:* Precision =  $TP / (TP + FP)$

TP (True Positives): Correct positive predictions

FP (False Positives): Incorrect positive predictions

- **Recall (Sensitivity):**

*Definition:* Measures the proportion of actual positives that were correctly identified.

*Formula:* Recall =  $TP / (TP + FN)$

FN (False Negatives): Missed positive predictions

- **Accuracy:**

*Definition:* Measures the proportion of all predictions that were correct.

*Formula:* Accuracy =  $(TP + TN) / (TP + FP + FN + TN)$

TN (True Negatives): Correct negative predictions

- **F1 Score:**

*Definition:* Harmonic mean of Precision and Recall, balancing the two metrics.

*Formula:* F1 Score =  $2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$

## BLEU: Bilingual Evaluation Understudy

BLEU is a method for evaluating the quality of text in machine-translation.

It works by comparing the machine-translated text to reference translations.

$$\log \text{BLEU} = \min \left( 1 - \frac{l_r}{l_c}, 0 \right) + \sum_{n=1}^N w_n \log p_n$$

$$p_n = \frac{\text{Number of ngrams in system and reference translations}}{\text{Number of ngrams in system translation}}$$

w = Weight for each n-gram (typically equal weight)

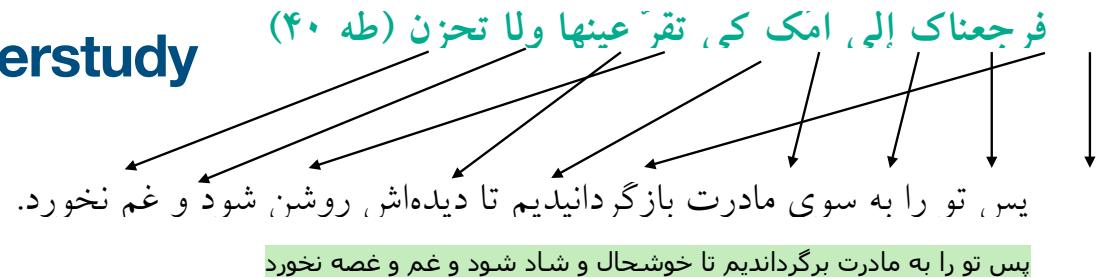
lc = length of hypothesis translation

lr = length of closest reference translation

Score Range: 0 to 1 (or 0 to 100%). Higher is better.

# Metrics

## BLEU: Bilingual Evaluation Understudy



BLEU is a method for evaluating the quality of text in machine-translation.

It works by comparing the machine-translated text to reference translations.

$$\log \text{BLEU} = \min \left( 1 - \frac{l_r}{l_c}, 0 \right) + \sum_{n=1}^N w_n \log p_n$$

$$p_n = \frac{\text{Number of ngrams in system and reference translations}}{\text{Number of ngrams in system translation}}$$

w = Weight for each n-gram (typically equal weight)

lc = length of hypothesis translation

lr = length of closest reference translation

Score Range: 0 to 1 (or 0 to 100%). Higher is better.

# Metric

## Recall-oriented Understudy for Gisting Evaluation (ROUGE)

Metric proposed to evaluate text summaries. It calculates recall score of the generated sentences corresponding to the reference sentences using n-grams.

$$ROUGE - N = \frac{\sum_{S \in R_{Sum}} \sum_{g_n \in S} C_m(g_n)}{\sum_{S \in R_{Sum}} \sum_{g_n \in S} C(g_n)}$$

$c_m$  represents the highest number of n-grams  
that are present in candidate as well as ground truth summaries  
 $R_{sum}$  reference summaries

ROUGE-L is based on the longest common subsequence (LCS)  
between our model output and reference:

R: The cat is on the mat.  
C: The cat and the dog.

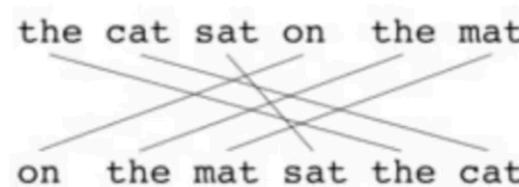
# Metrics

## Metric for Evaluation of Translation with Explicit Ordering (METEOR)

- METEOR is another metric for evaluating machine translation:
  - Considering synonyms, stemming, and paraphrasing
  - More weights to Recall
  - Better correlates with human judgement

$$F_{\text{mean}} = \frac{10PR}{R + 9P} \quad M = F_{\text{mean}} (1 - \text{penalty})$$

- Penalty of ordering



$$\text{Penalty} = 0.5 \times \left( \frac{\text{number of chunks}}{\text{number of unigrams matched}} \right)^3$$

- "number of chunks" refers to the count of non-contiguous sequences of matched words in the candidate translation.
- "number of unigrams matched" is the total count of matched unigrams (individual words) in the candidate translation.

# Metrics

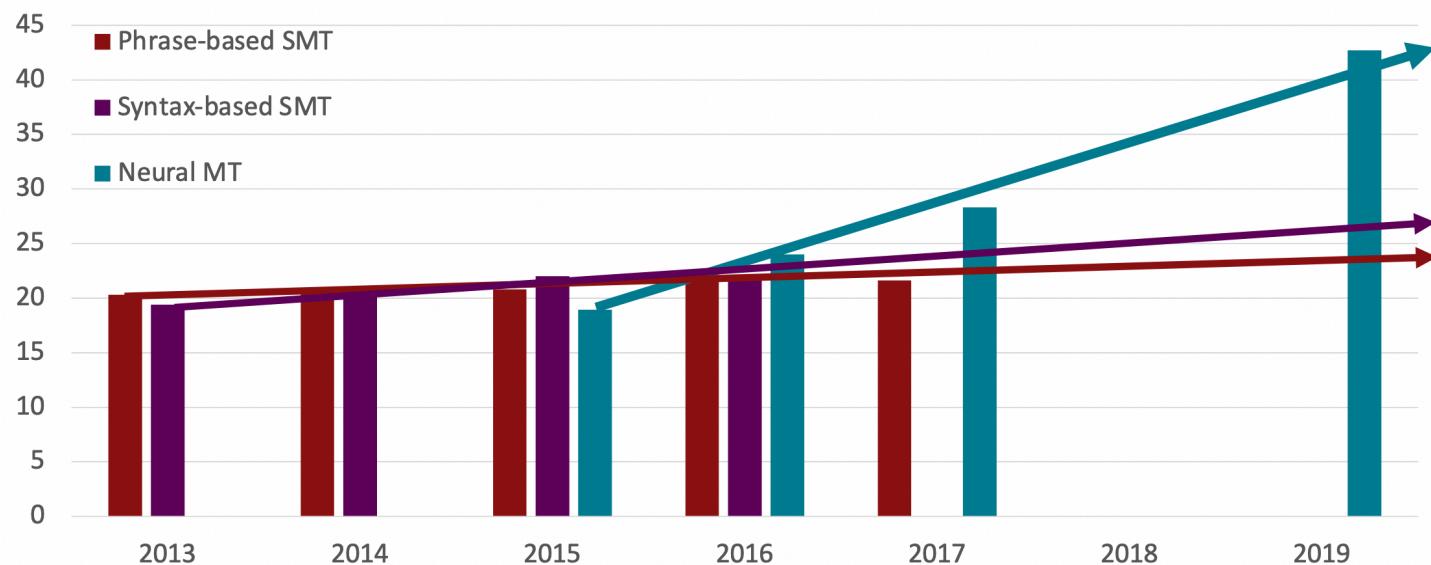
## MRR (Mean Reciprocal Rank)

MRR is a statistic measure for evaluating any process that produces a list of possible responses to a sample of queries, ordered by probability of correctness.

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

Q's are queries.

# پیشرفت ترجمه ماشینی



- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

# چالشها

- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

- کلمات کم تکرار

- حوزه‌های متفاوت

- زبانهای low-resource

- تشخیص ناصحیح ضمیر

- متنهای طولانی

# ترجمه ماشینی آماری

یادگیری یک مدل آماری از دیتا بگونه‌ای که بتوانیم بهترین ترجمه جمله  $X$  در زبان مبداء | را در زبان مقصد  $I$ ,  $(y)$  بیابیم.

$$\arg \max_{y \in \mathcal{Y}} P(y|x)$$

$$= \arg \max_{y \in \mathcal{Y}} P(y)P(x|y)$$

مدل ترجمه کلمات و عبارات      مدل زبانی در زبان مقصد  
یادگیری از داده موازی      یادگیری از داده تک زبانه

- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

## يادگيري تطبيق لغت (alignment)

$$\arg \max_{y \in \mathcal{Y}} P(y) P(x|y)$$

$p(x | y) = \sum_{\mathbf{a} \in \mathcal{A}} p(\mathbf{a} | y) p(x | \mathbf{a}, y)$

فرجعنای ای امک کی تقر عینها ولآ تحزن (له ۴۰)  
پس تو را به سوی مادرت بازگردانیدیم تا دیده اش روشن شود و غم نخورد.

- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

# يادگیری تطبیق لغت (alignment)

$$\arg \max_{y \in \mathcal{Y}} P(y) P(x|y)$$

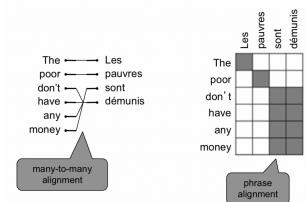
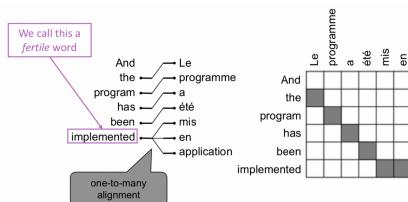
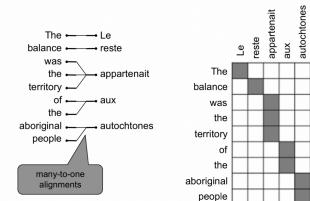
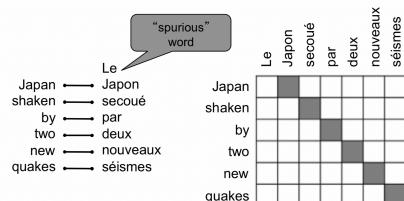
$p(x | y) = \sum_{\mathbf{a} \in \mathcal{A}} p(\mathbf{a} | y) p(x | \mathbf{a}, y)$

فرجهنگ ای امک کی تقر عینها ولاتخزن (طه ۴۰)

پس تو را به سوی مادرت بازگردانیدم تا دیده اش روشن شود و غم نخورد.

پس تو را به مادرت برگردانیدم تا خوشحال و شاد شود و غم و غصه نخورد

- تفاوت های نوع شناسی باعث پیچیده شدن گراف تطبیق می شود.



Brown, Peter F. (1993). "The mathematics of statistical machine translation: Parameter estimation", CL.

- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

# IBM Alignment models

Model 1	Independent word translation (order doesn't matter)
Model 2	Word translation + distance between source and target position
Model 3	Word translation + fertility (how many target words a source word can align to)
Model 4	Word translation + relative ordering among target words of same source
Model 5	(Fixes deficiency of model 4)
HMM (Vogel et al. 1996)	Word translation plus relative ordering

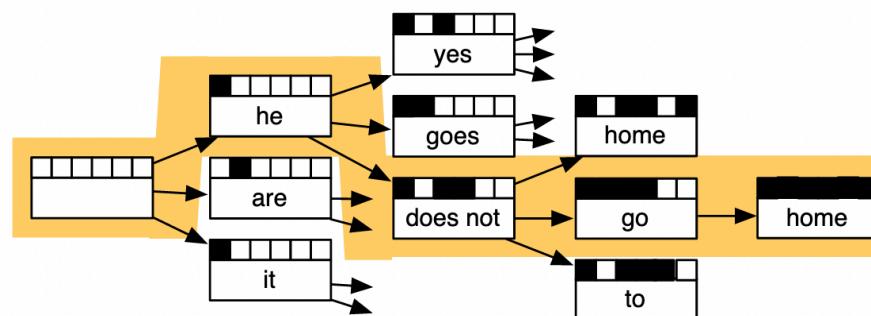
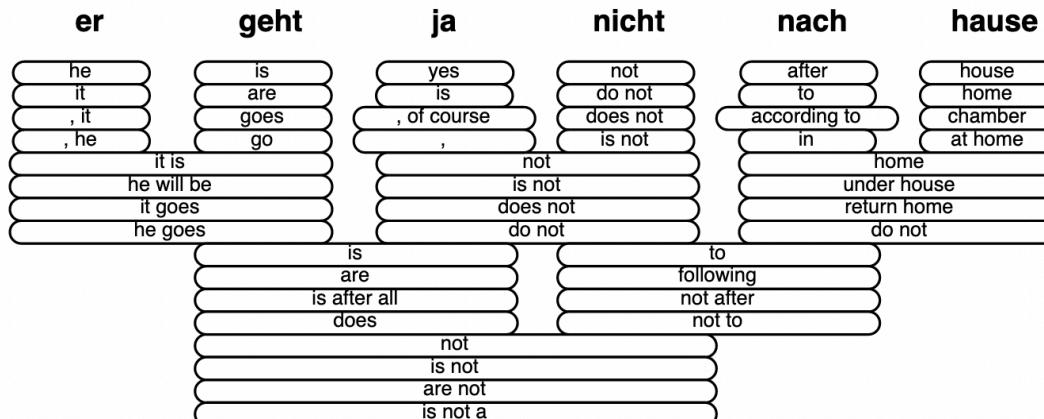
Brown, Peter F. (1993). "The mathematics of statistical machine translation: Parameter estimation", CL.

- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

# دیکدینگ ترجمه ماشینی آماری

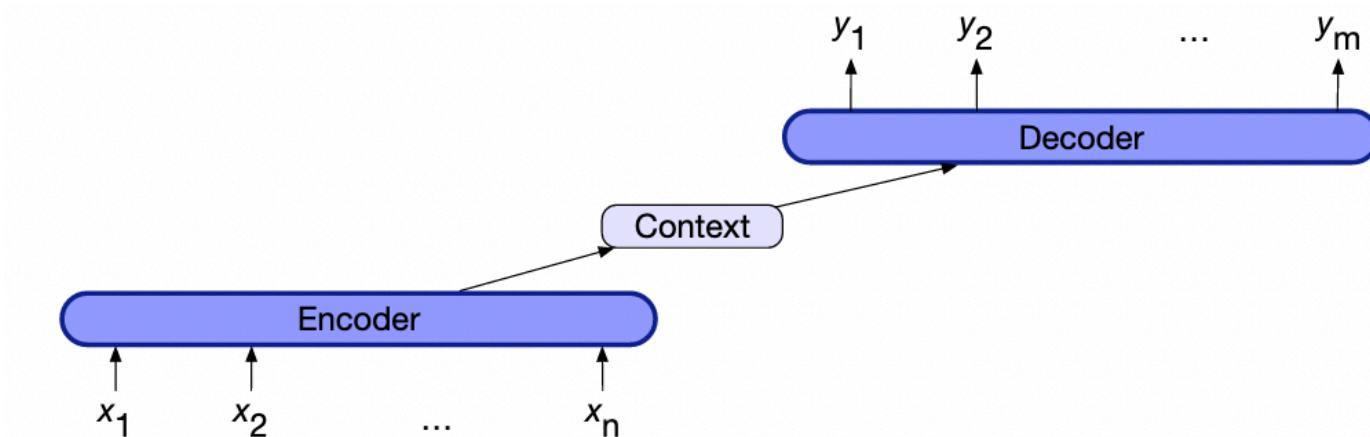
$$\arg \max_{y \in \mathcal{Y}} P(y) P(x|y)$$

مدل ترجمه کلمات و عبارات      مدل زبانی در زبان مقصد  
یادگیری از داده موازی      یادگیری از داده تک زبانه



# ترجمه مبتنی بر شبکه عصبی

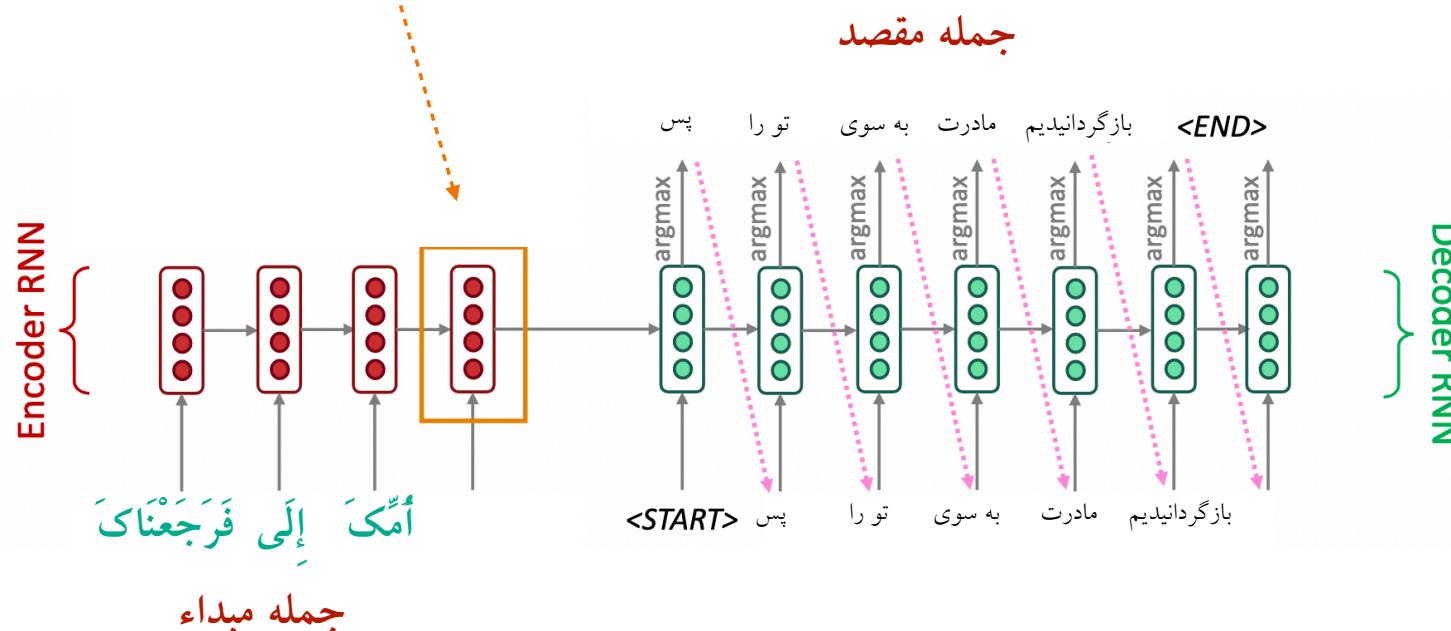
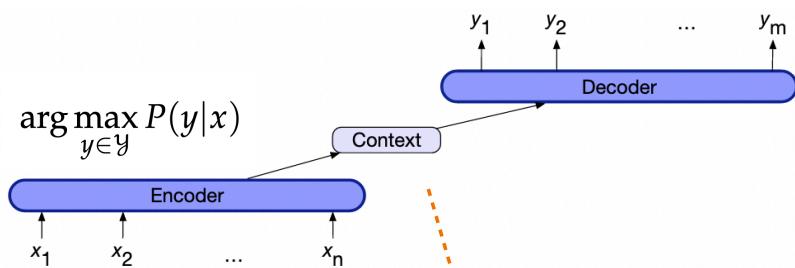
$$\arg \max_{y \in \mathcal{Y}} P(y|x)$$



- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

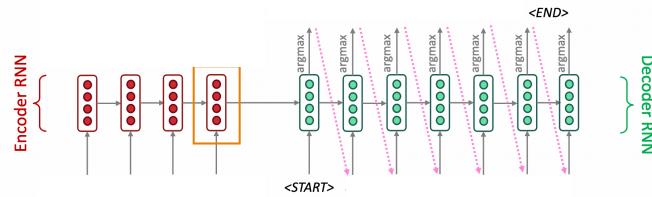
## مدل seq-to-seq



Sutskever et al. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems*. 2014.

- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

# کاربردها: مدل seq-to-seq



- خلاصه‌سازی متن

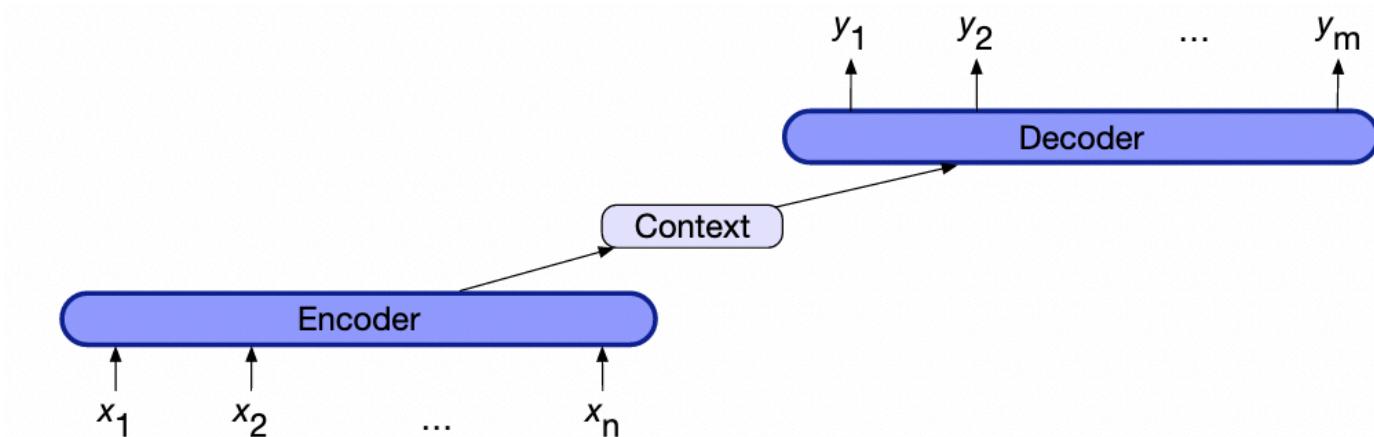
- دیالوگ

- تولید کد

- شکست لغات به morpheme‌ها

# ترجمه مبتنی بر شبکه عصبی

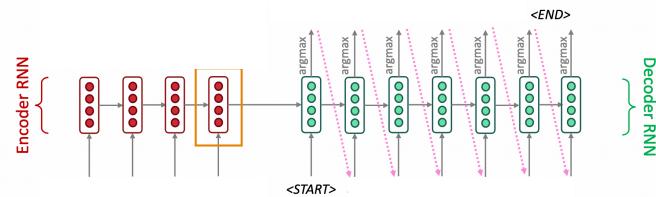
$$\arg \max_{y \in \mathcal{Y}} P(y|x)$$



- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

# مدل seq-to-seq



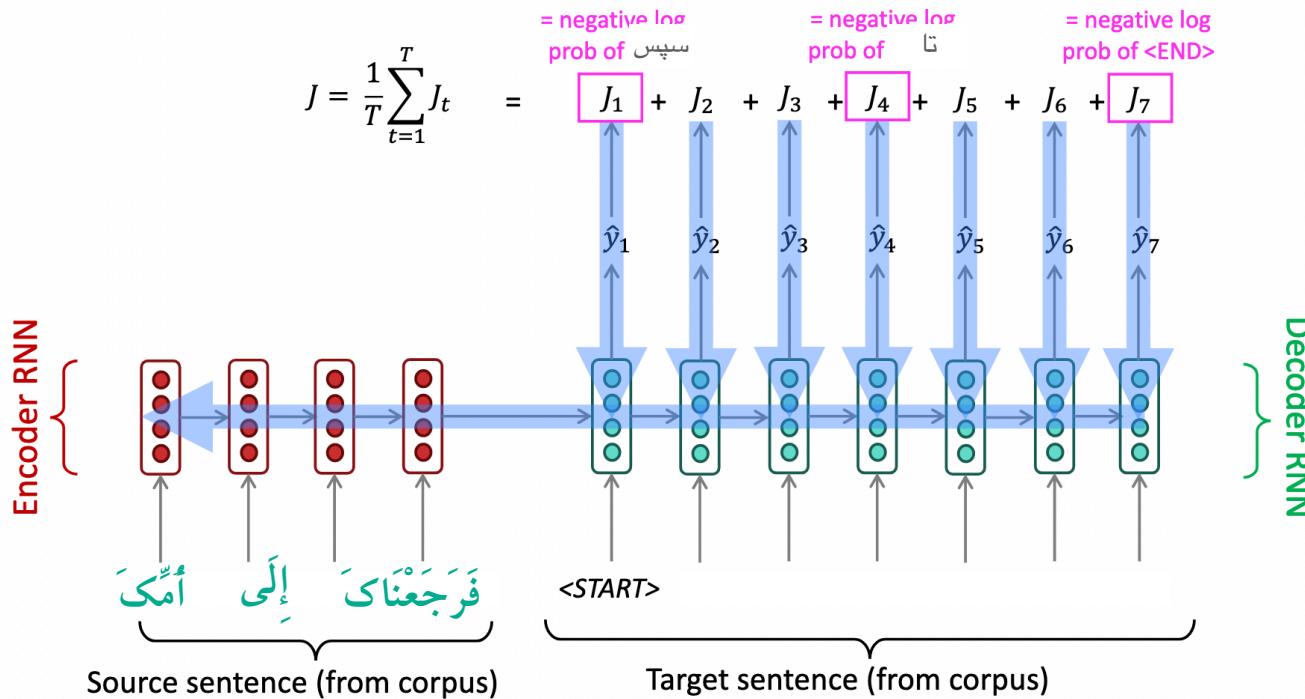
مدل زبانی شرطی

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

چگونگی یادگیری؟

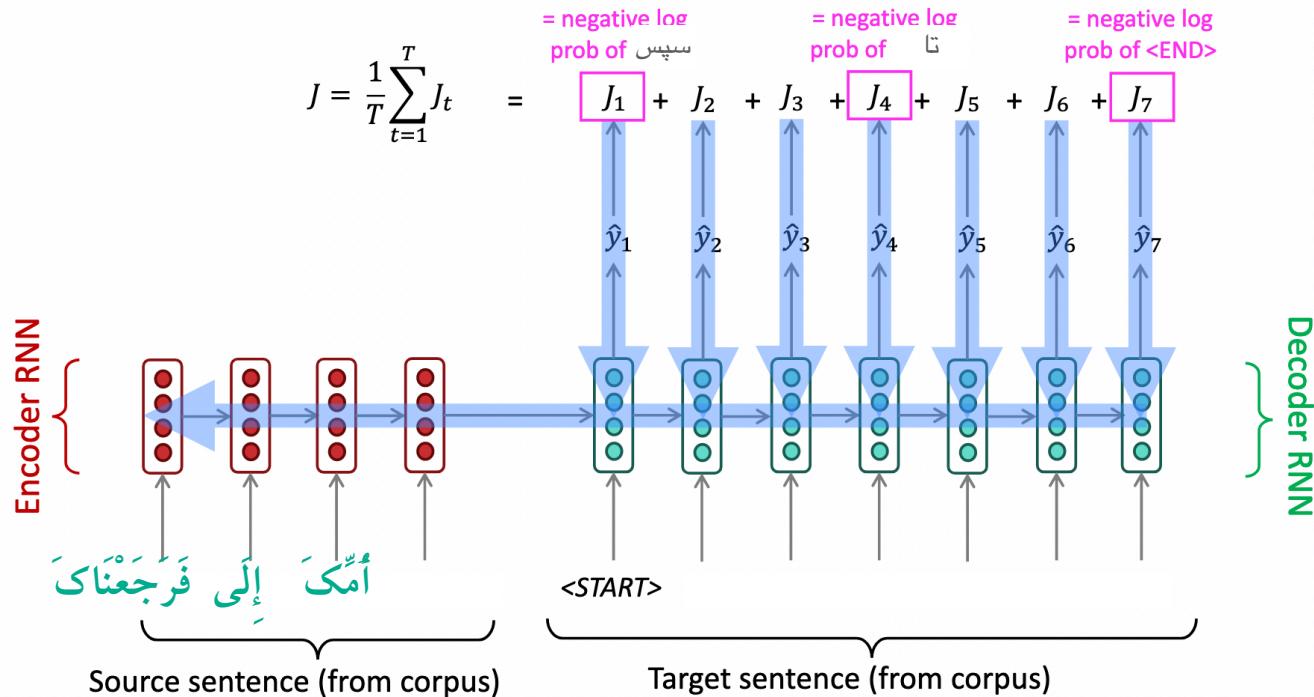
- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

# یادگیری end-to-end مدل seq-to-seq

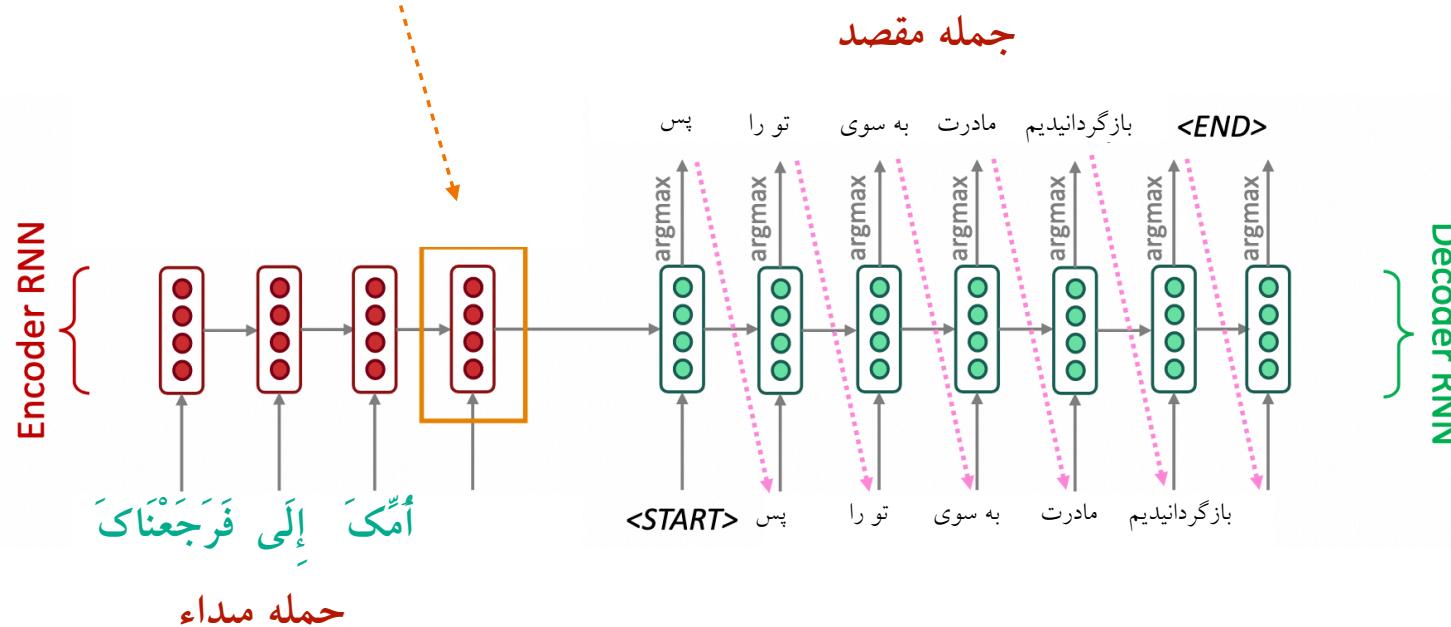
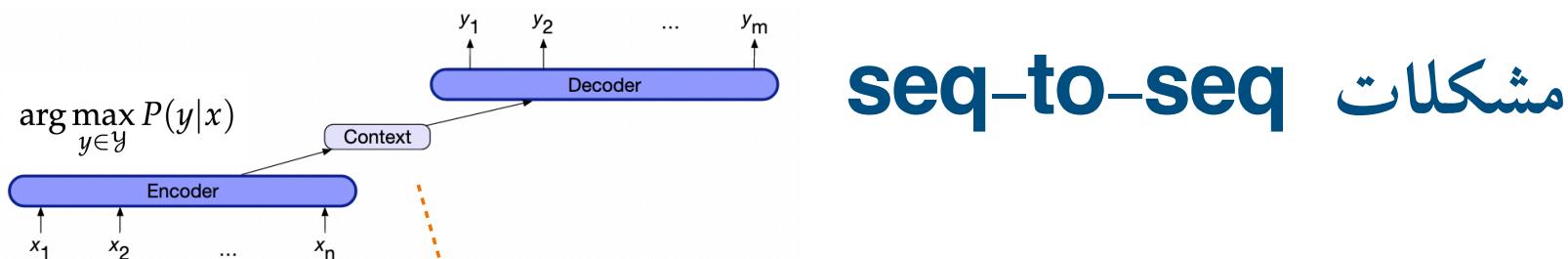


- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

# یادگیری end-to-end مدل seq-to-seq



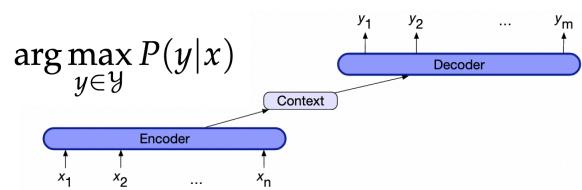
- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال



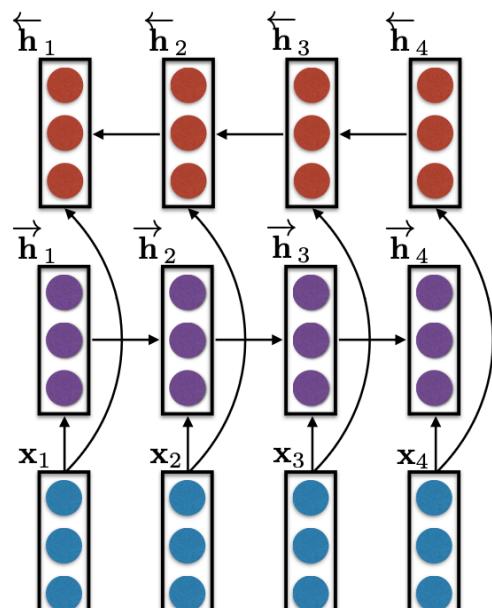
Sutskever et al. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems*. 2014.

- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

# Attentive seq-to-seq



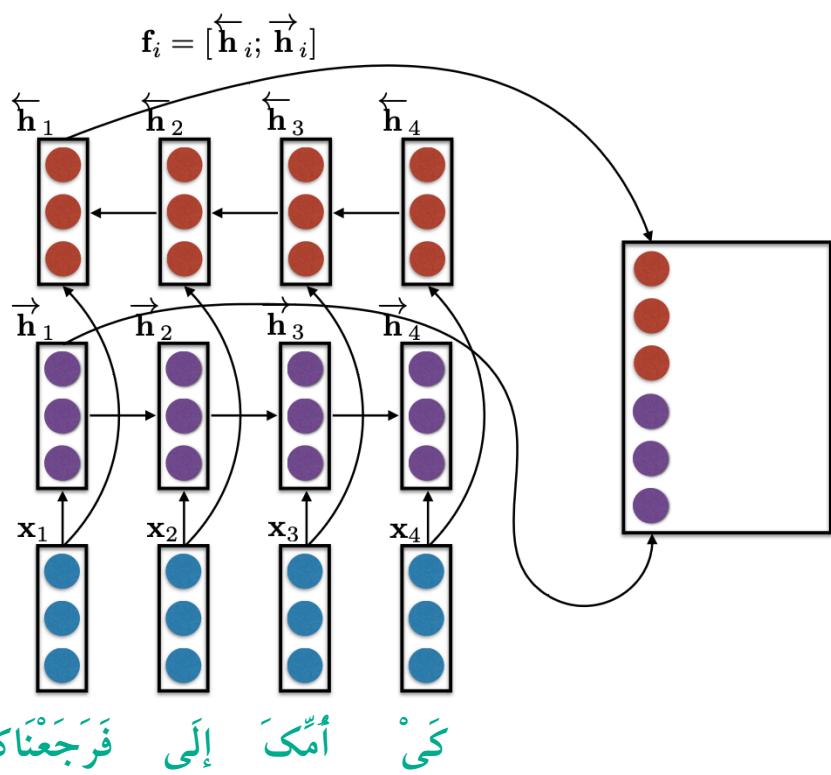
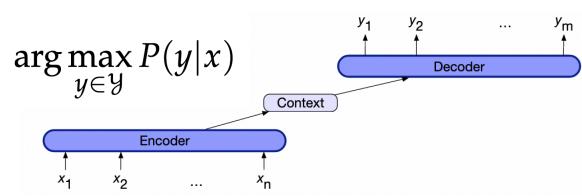
$$\mathbf{f}_i = [\overleftarrow{\mathbf{h}}_i; \overrightarrow{\mathbf{h}}_i]$$



كَيْ أُمْكِنْ إِلَى فَرَجَعَنَاكَ

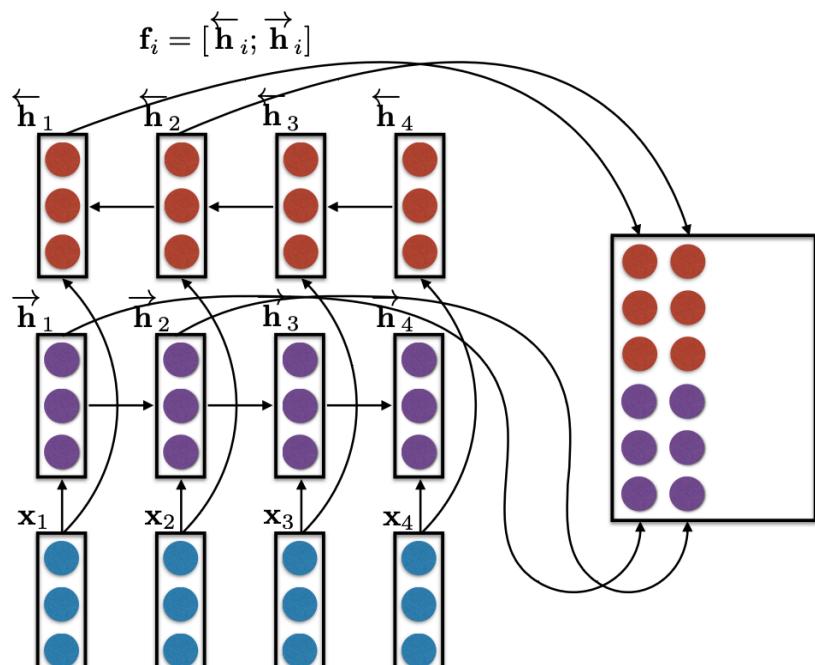
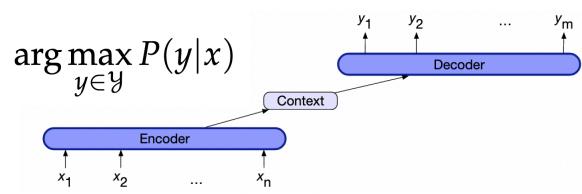
- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

# Attentive seq-to-seq



- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

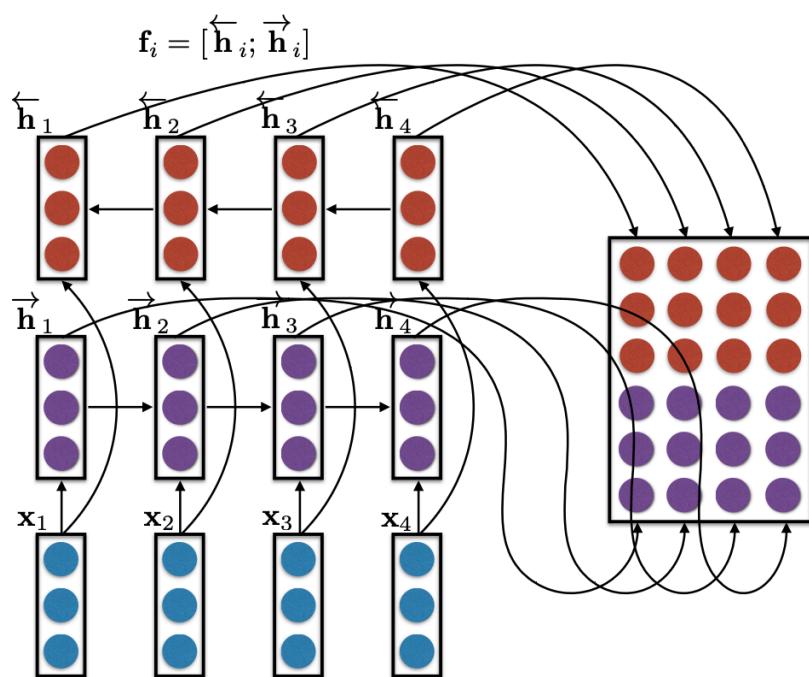
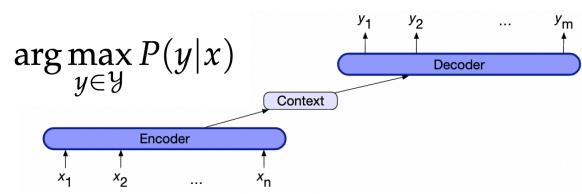
# Attentive seq-to-seq



كىْ أمّكَ إِلى فَرَجَعَنَاكَ

- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

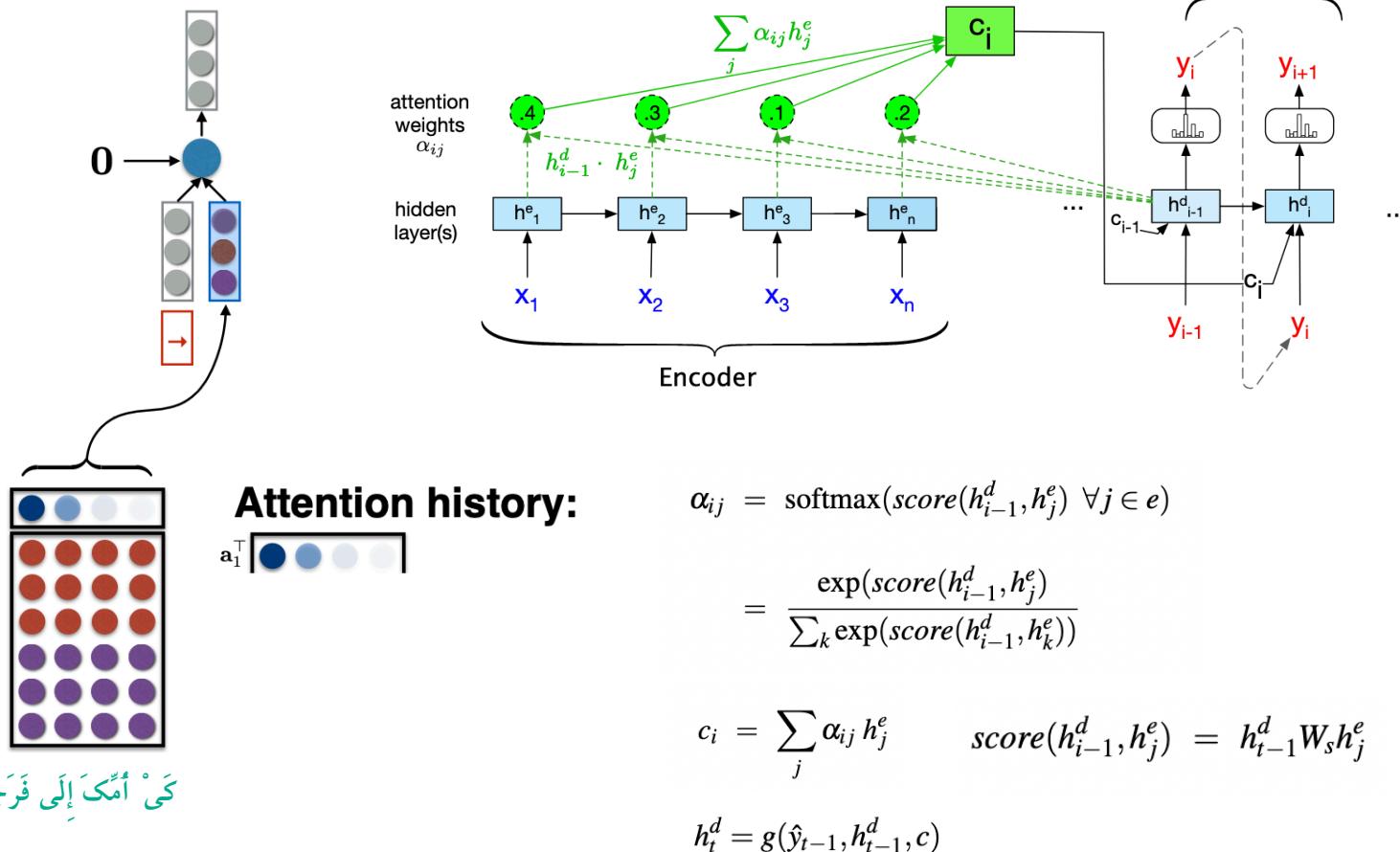
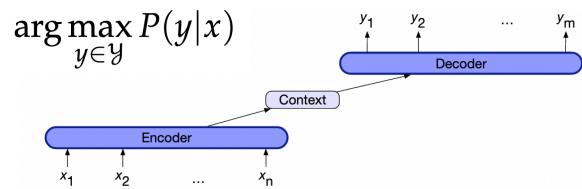
# Attentive seq-to-seq



كىنْ امّكِ إِلى فَرَجَعَناكَ

- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

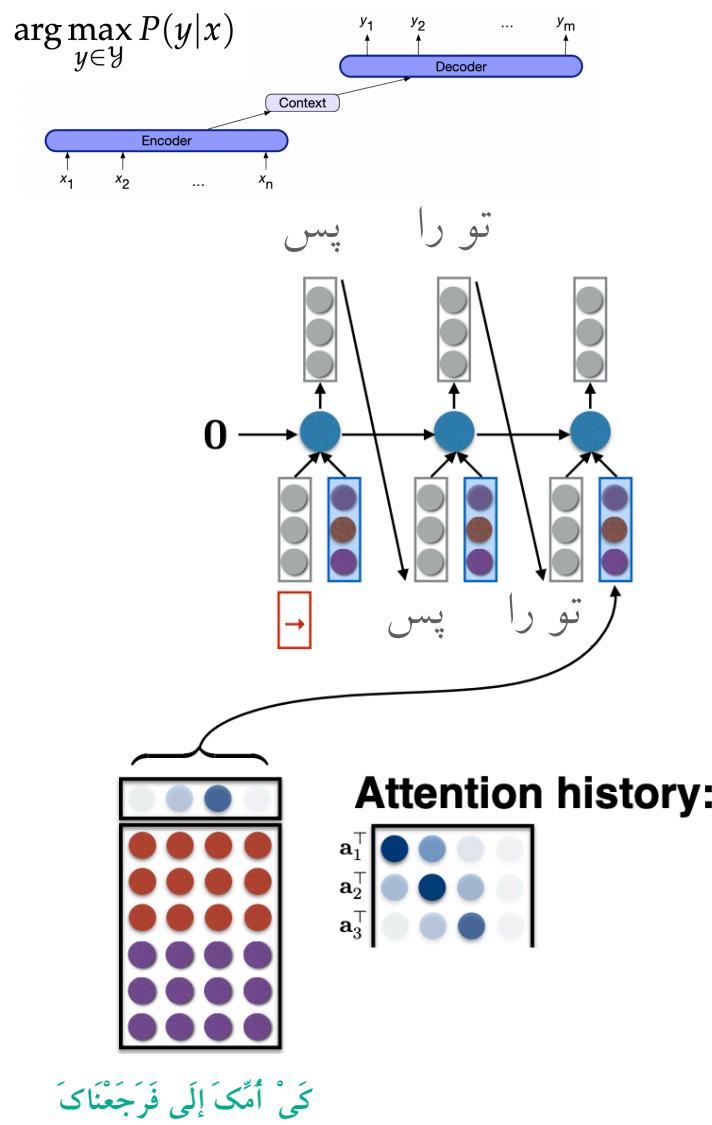
# Attentive seq-to-seq



كىْ امكَ إِلى فَرَجَعَنَا

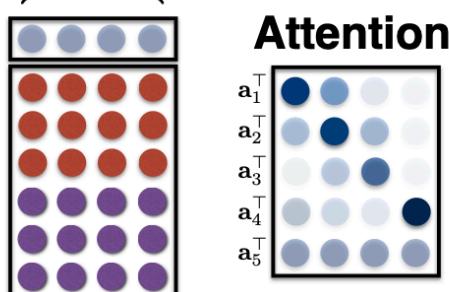
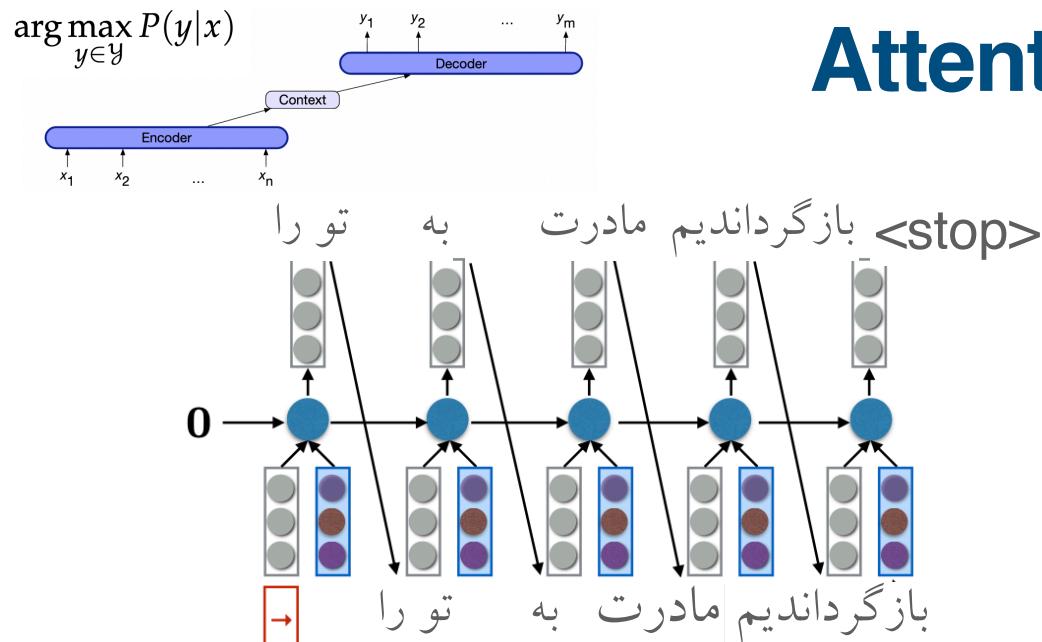
- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

# Attentive seq-to-seq



- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

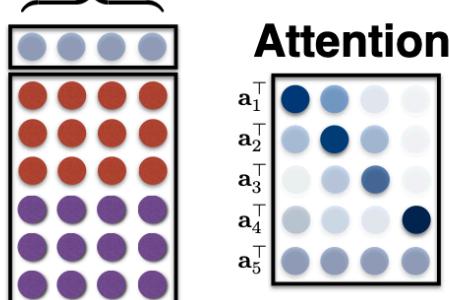
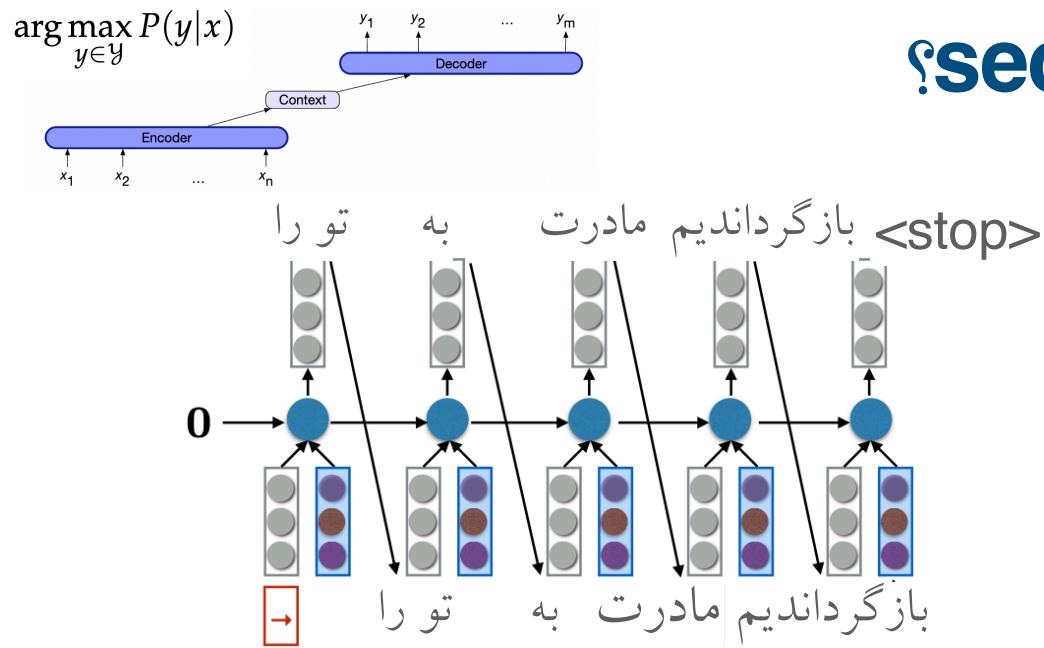
# Attentive seq-to-seq



كى امك إلى فرجتنا!

- مقدمات ترجمه ماشینی
- ترجمه ماشینی آماری
- ترجمه ماشینی نورال

## مشکلات seq-to-seq



کی اُمکِی فرجعتاً

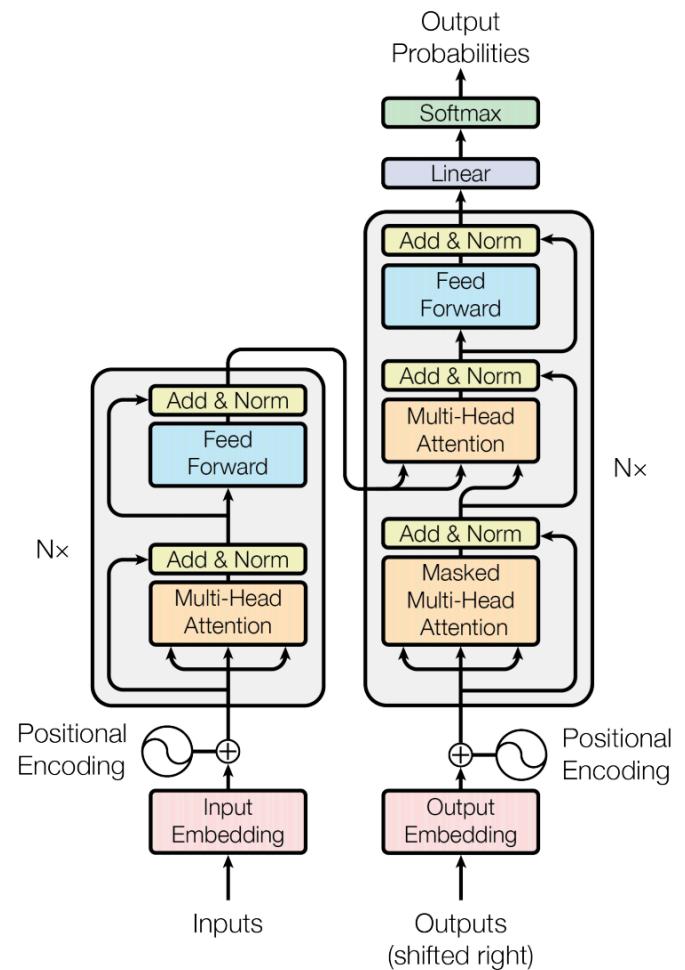
# Earliest works

- Introducing transformers
- Multihead attention
- For machine translation

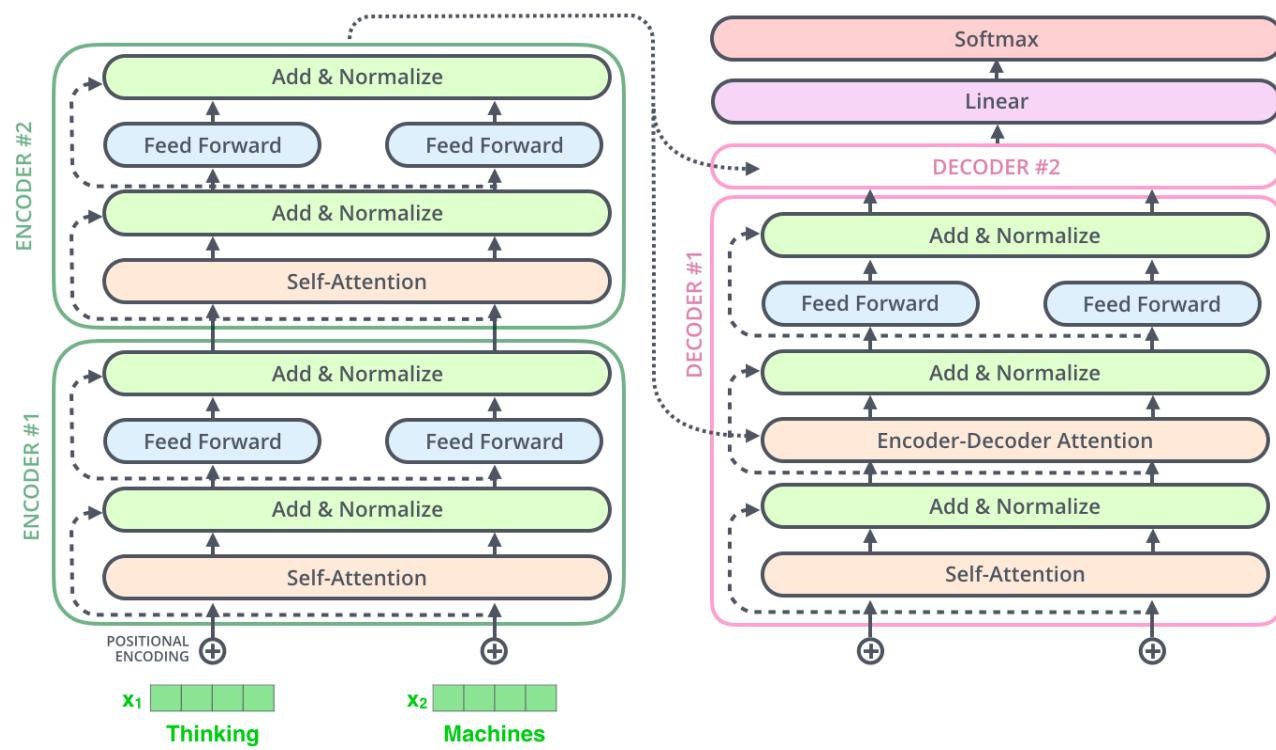
$$\Pr(y_1, \dots, y_n | \mathbf{x}) = \prod_i^n \Pr(y_i | y_{i-1}, \dots, y_1, \mathbf{x})$$

- Target seq  $\mathbf{y} = (y_1, \dots, y_{T_y})$
- Source seq  $\mathbf{x} = (x_1, \dots, x_{T_x})$

Ashish Vaswani *et al.*  
**Attention is All you Need.** NIPS 2017: 5998-6008.

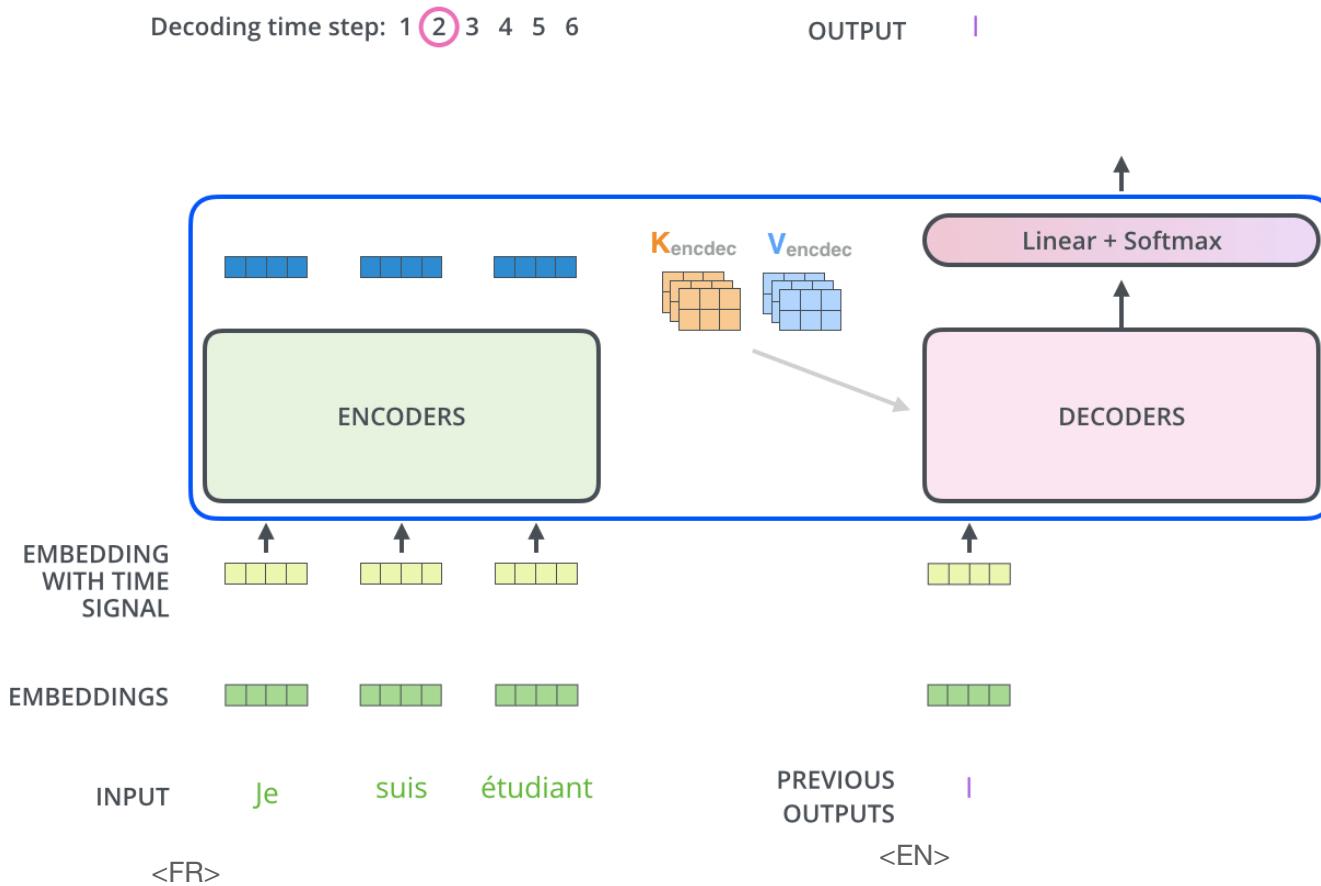


# Translation Model



Reference: <https://jalammar.github.io/>

# Animation of the Translation Model



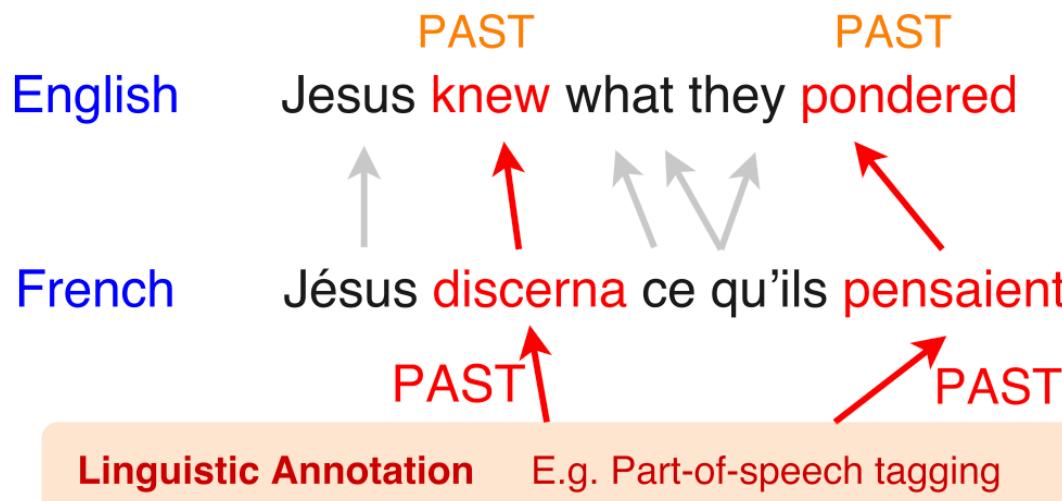
Reference: <https://jalammar.github.io/>

# Annotation Project , Alignment

کاربرد دیگر تطبیق کلمات

در تولید داده

**Important area of NLP research:** Yarowsky et al. (2001); Spreyer and Frank (2008); Padó & Lapata (2009); Das and Petrov (2011); Agić et al. (2016).



# منبع تحلیل احساس برای ۱۰۰۰ زبان

