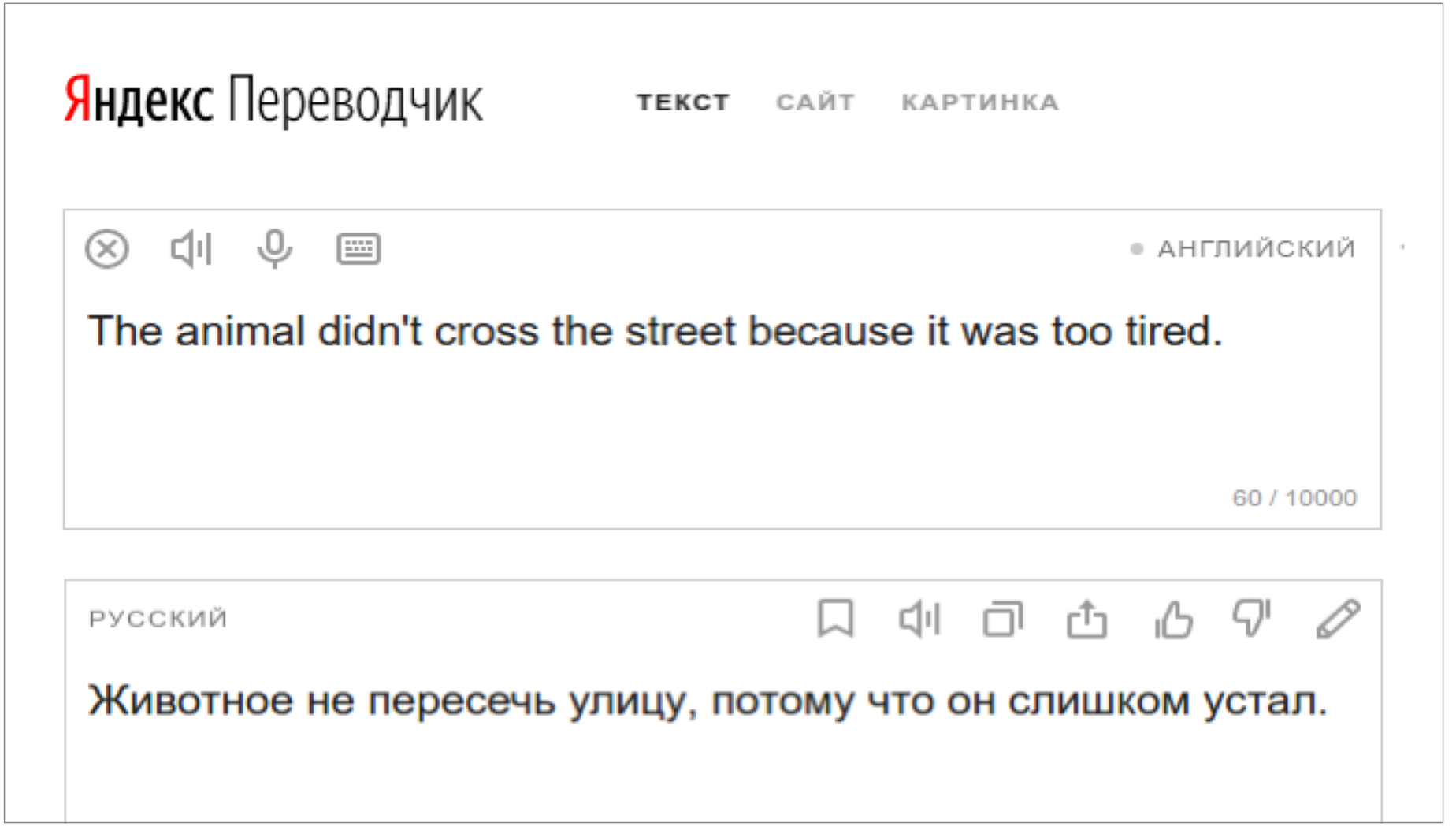
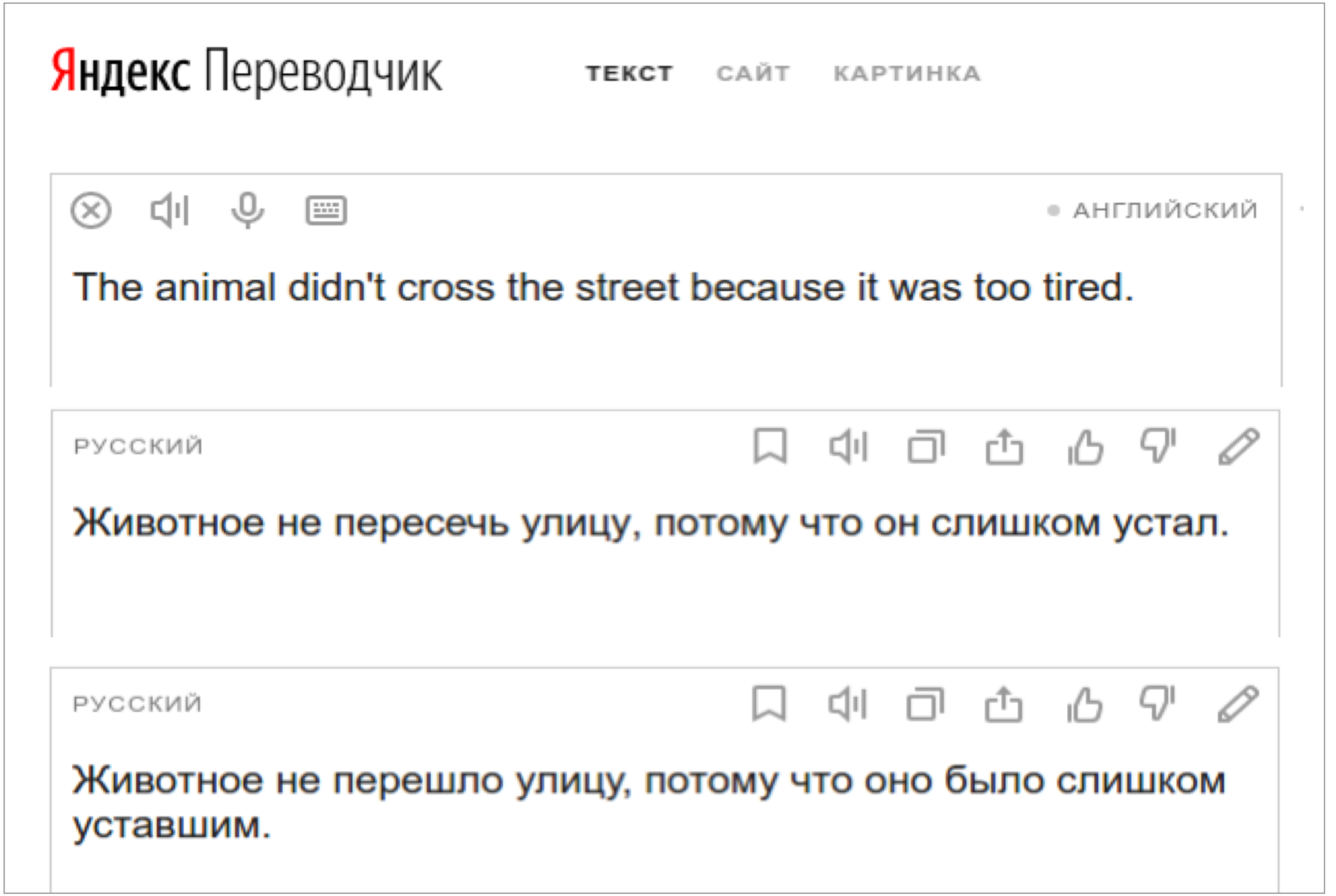
Yandex Translate

Neural Machine Translation

David Talbot


Яндекс Переводчик
ТЕКСТ
САЙТ
КАРТИНКА
АНГЛИЙСКИЙ
The animal didn't cross the street because it was too tired.
60 / 10000


Яндекс Переводчик
ТЕКСТ
САЙТ
КАРТИНКА
АНГЛИЙСКИЙ
The animal didn't cross the street because it was too tired.
60 / 10000
РУССКИЙ
Животное не пересечь улицу, потому что он слишком устал.


Яндекс Переводчик
ТЕКСТ
САЙТ
КАРТИНКА
АНГЛИЙСКИЙ
The animal didn't cross the street because it was too tired.
РУССКИЙ
Животное не пересечь улицу, потому что он слишком устал.
РУССКИЙ
Животное не перешло улицу, потому что оно было слишком уставшим.

Problems with Phrase-based MT

- Unrealistic independence assumptions
- Curse of dimensionality
- Impossible to optimize end-to-end
- Reliance on hand-crafted language specific features

Neural Machine Translation

› Direct model of conditional translation probability

$$\sout{e^* = argmax_f \Pr(e) \Pr(f|e)}$$

$$e^* = argmax_e \Pr(e|f)$$

Neural Machine Translation

› Direct model of conditional probability of translation given source

$$\sout{e^* = argmax_f \Pr(e) \Pr(f|e)}$$

$$e^* = argmax_e \Pr(e|f)$$

$$= \prod_{i=1}^{I} \Pr(e_i | e_{j<i}, f)$$

Neural Machine Translation

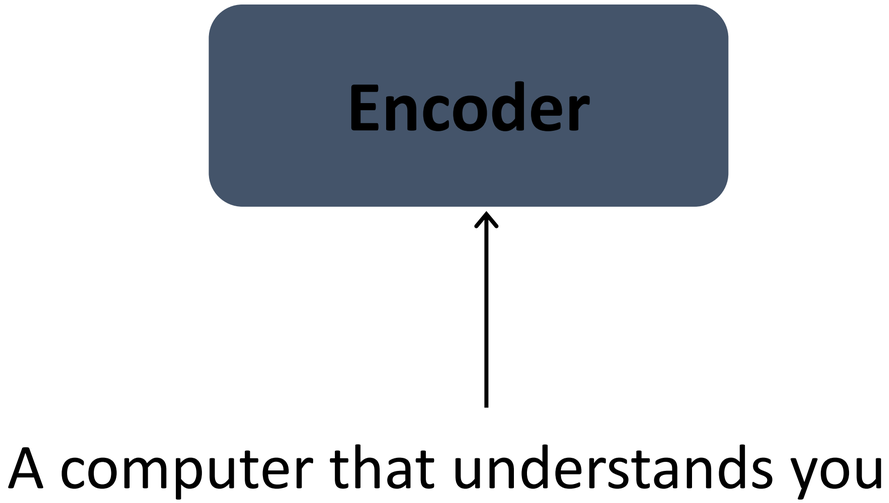
› Direct model of conditional probability of translation given source

$$\sout{e^* = argmax_f \Pr(e) \Pr(f|e)}$$

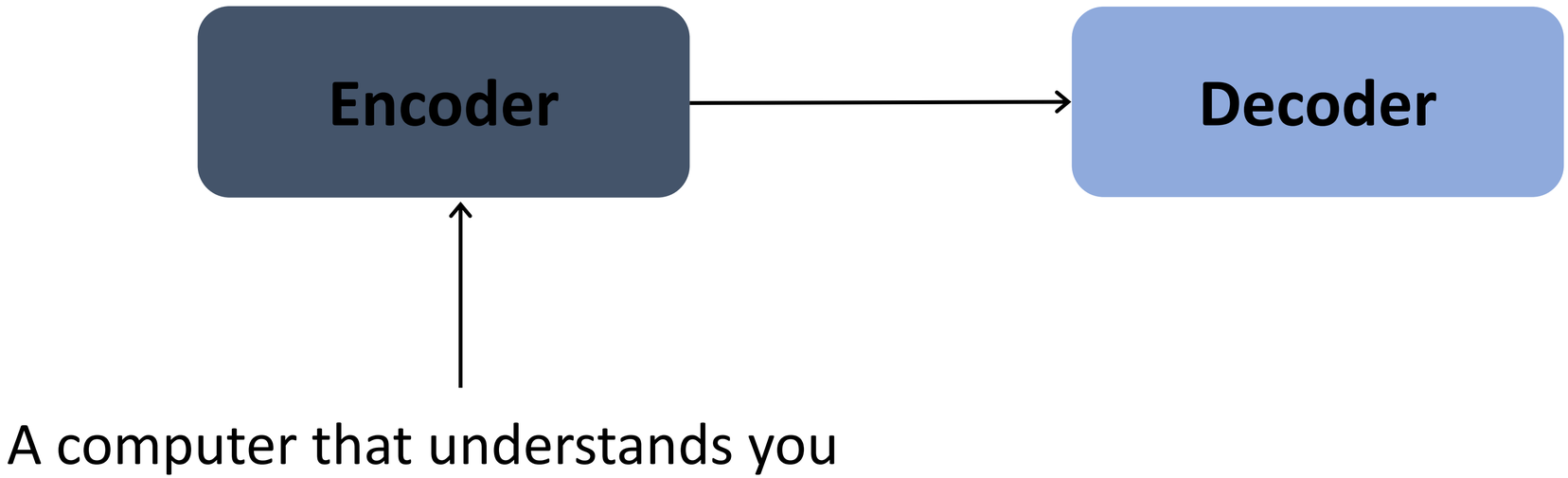
$$e^* = argmax_e \Pr(e|f)$$

$$= \prod_{i=1}^{I} \Pr(e_i | e_{j<i}, f)$$

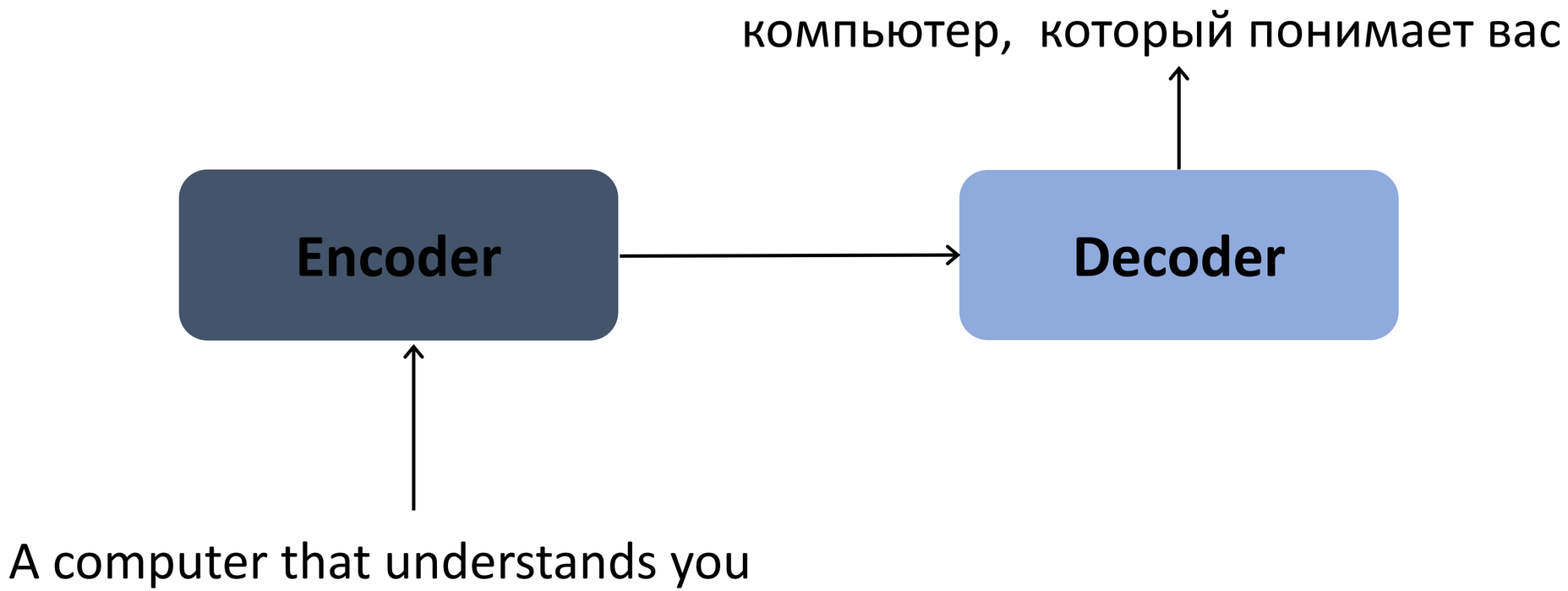
Neural Machine Translation



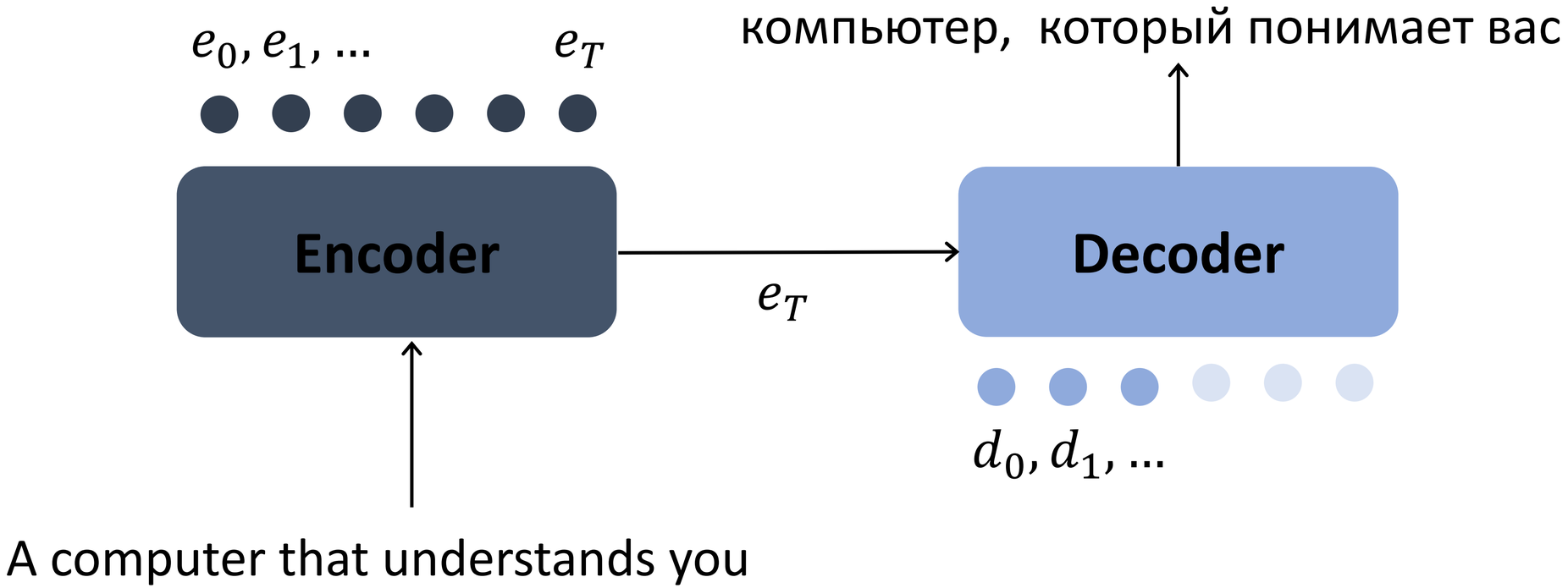
Neural Machine Translation



Neural Machine Translation



Neural Machine Translation



Encoder Decoder

- No explicit independence assumptions
- Learned representations (word embeddings)
- End-to-end optimization

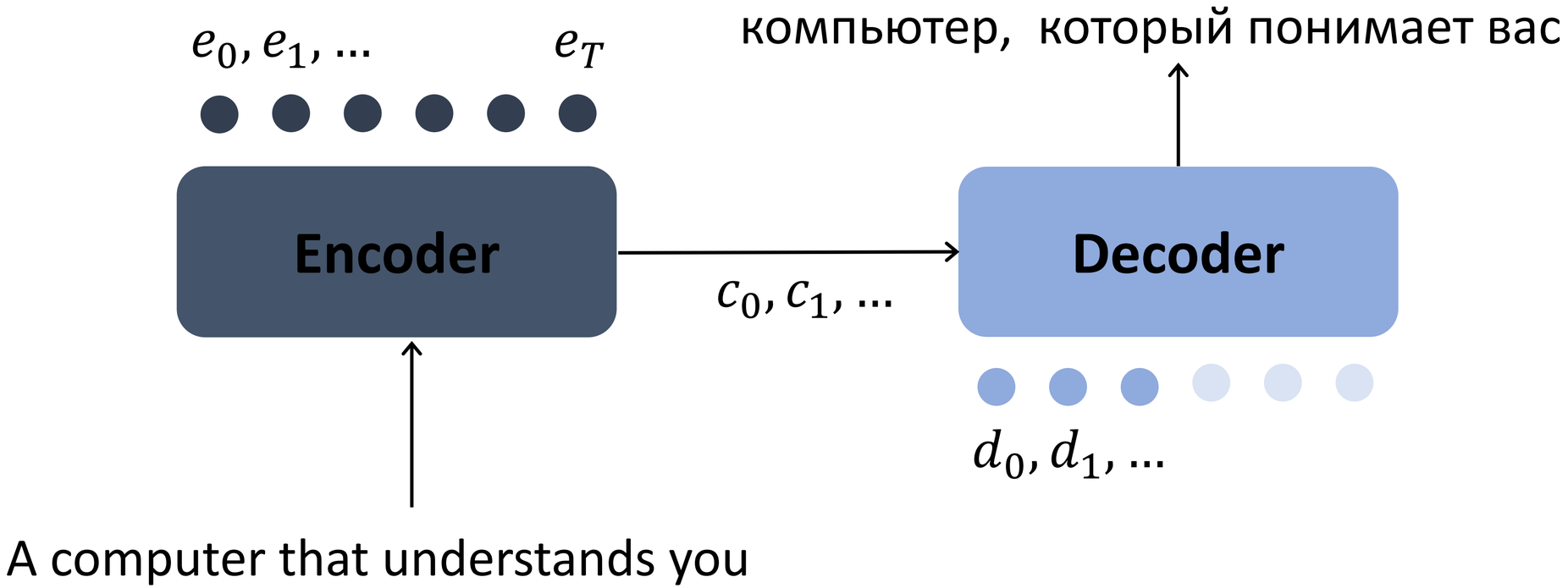
Problems with Vanilla Encoder Decoder

- Poor performance on long sentences
- Bias towards shorter candidates (?)
- Fluent but inadequate output
- No guarantee that all input words are translated

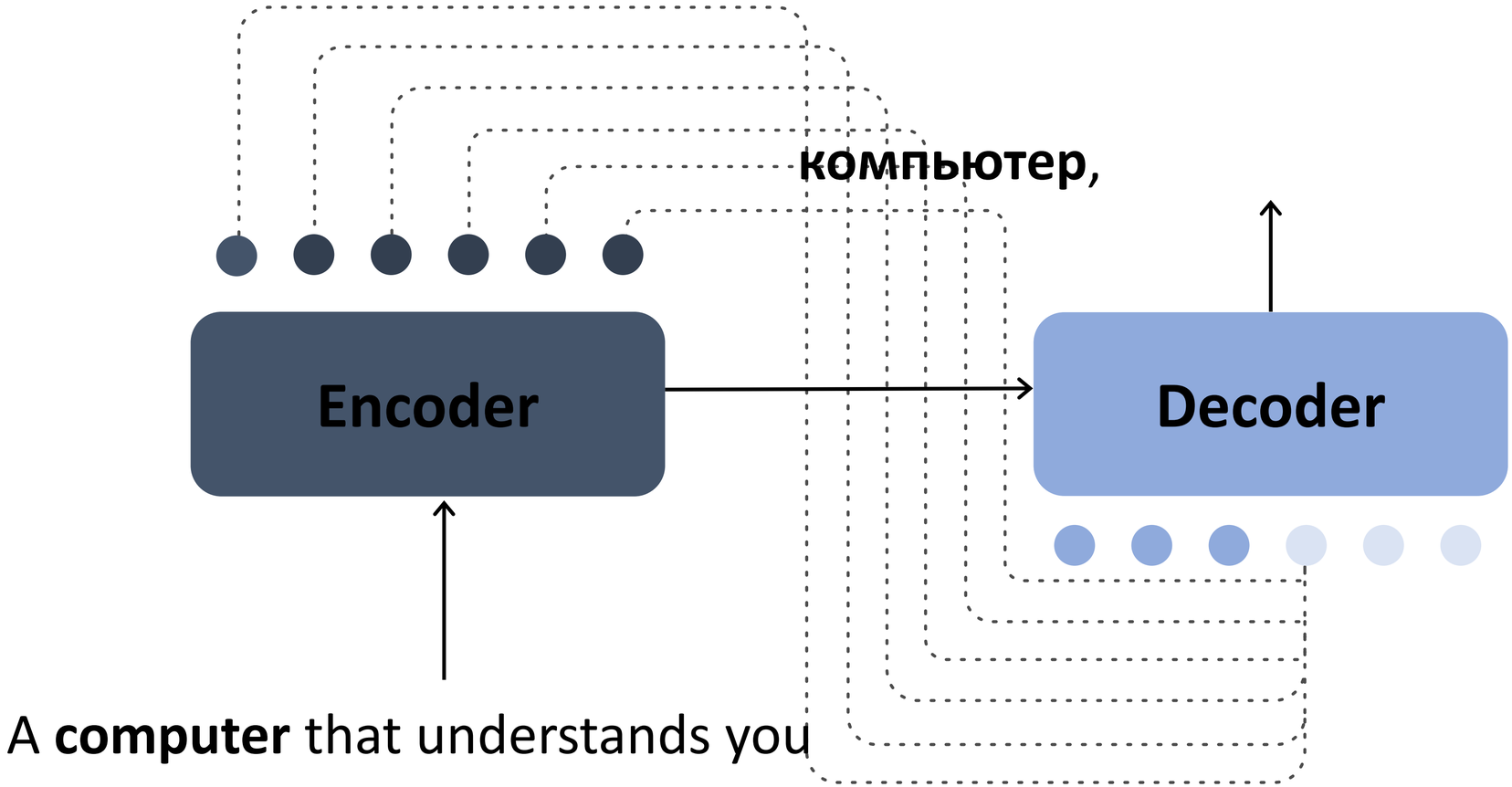
Attention

- Allow model to scale with sentence length
- Allow model to focus on specific information during translation
- Not quite word alignments

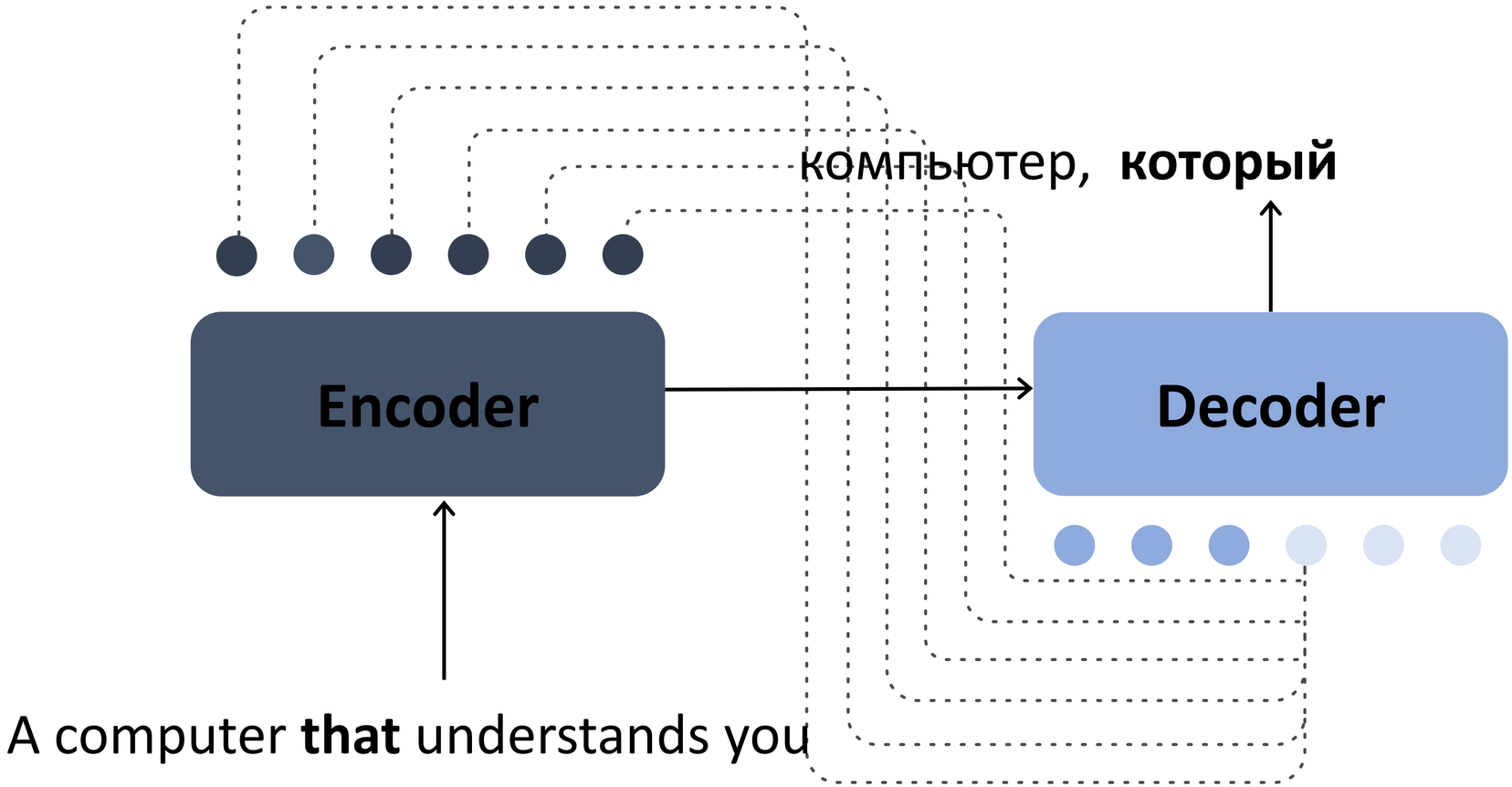
Neural Machine Translation



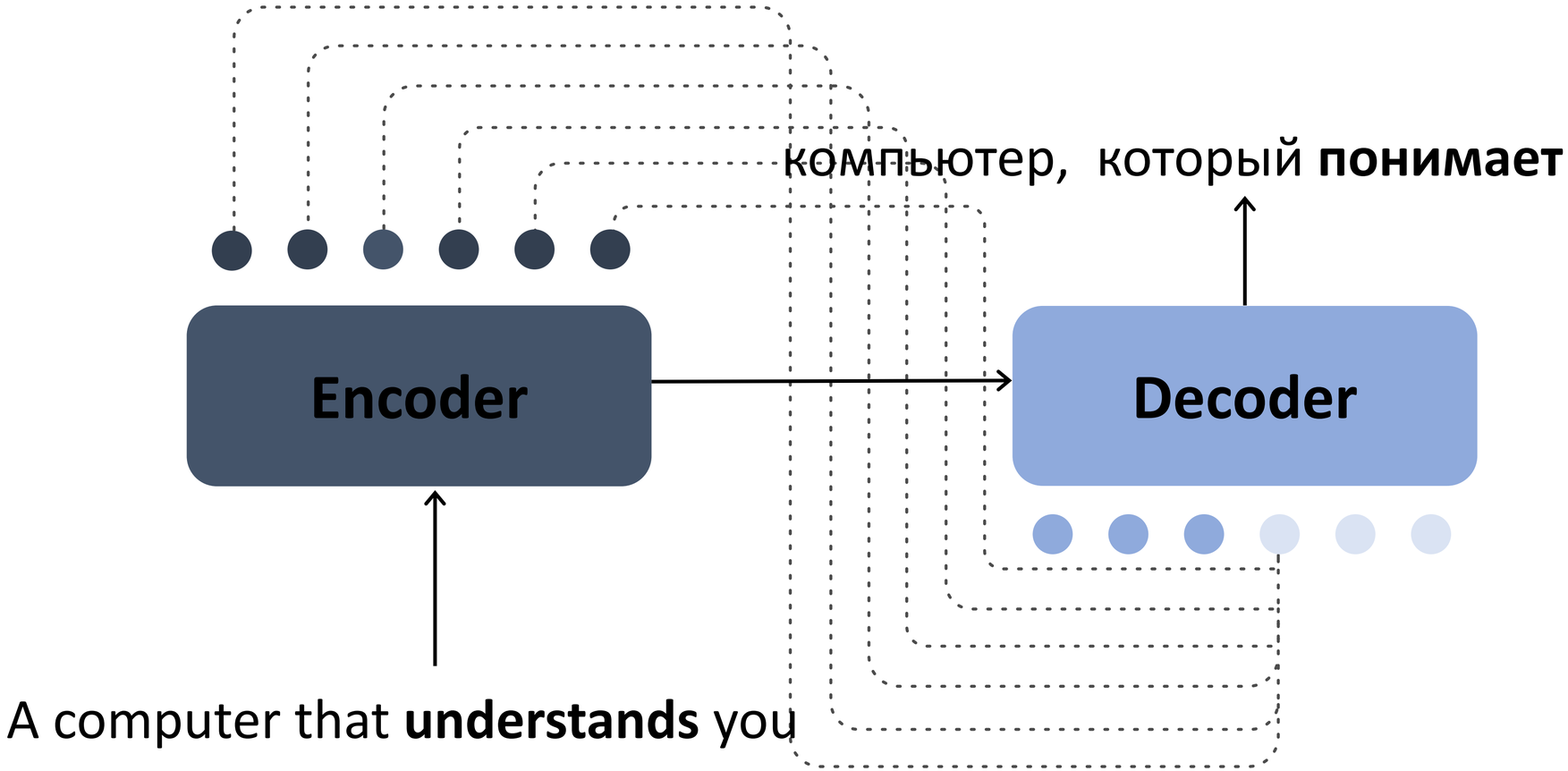
Neural Machine Translation



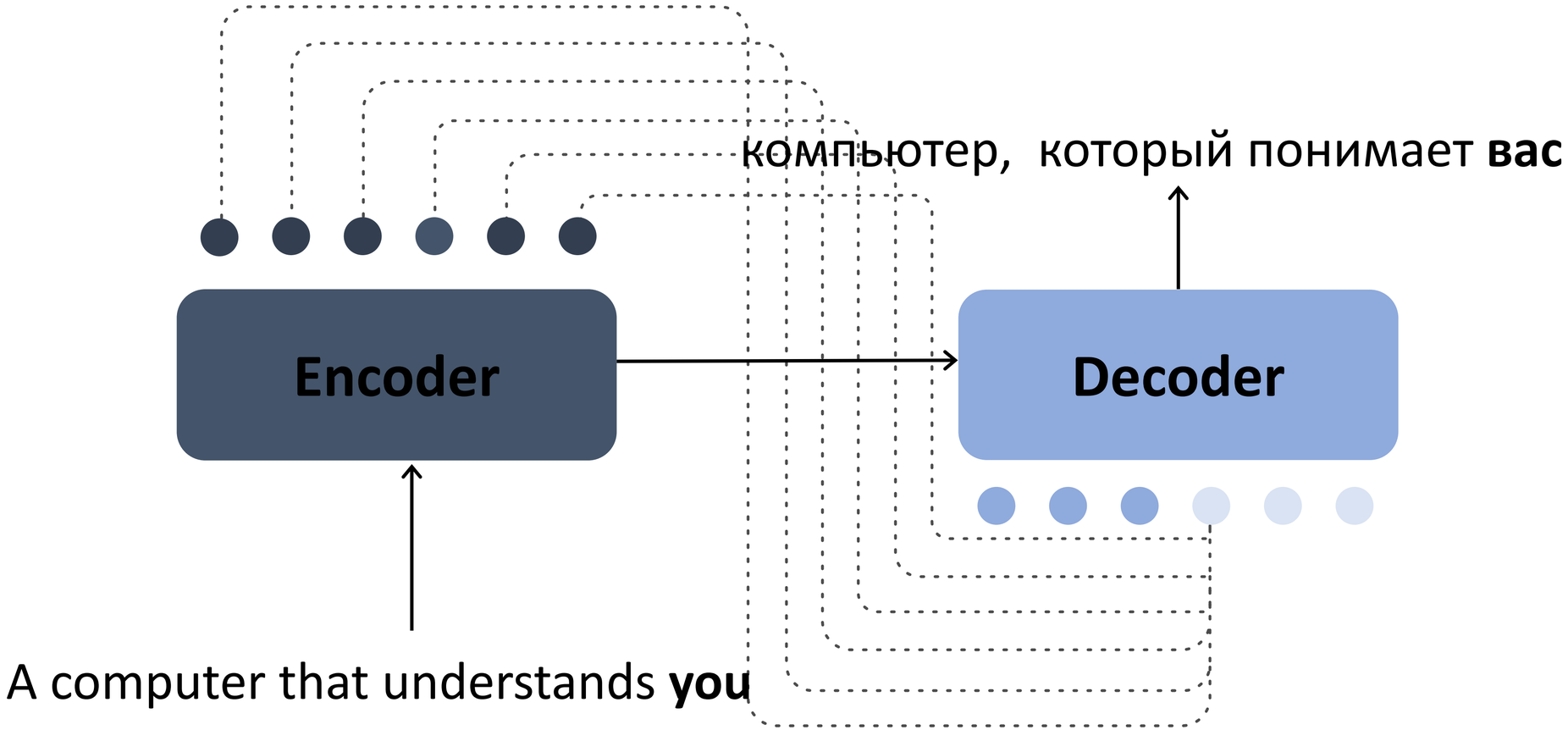
Neural Machine Translation



Neural Machine Translation

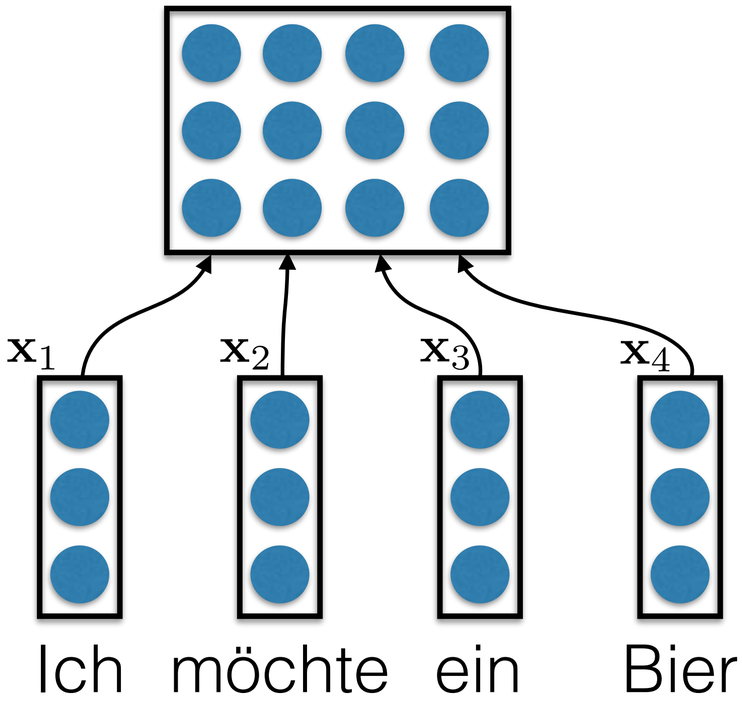


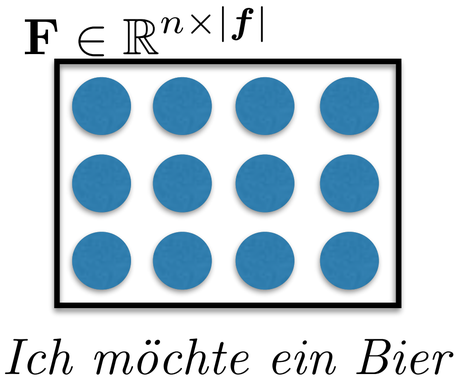
Neural Machine Translation

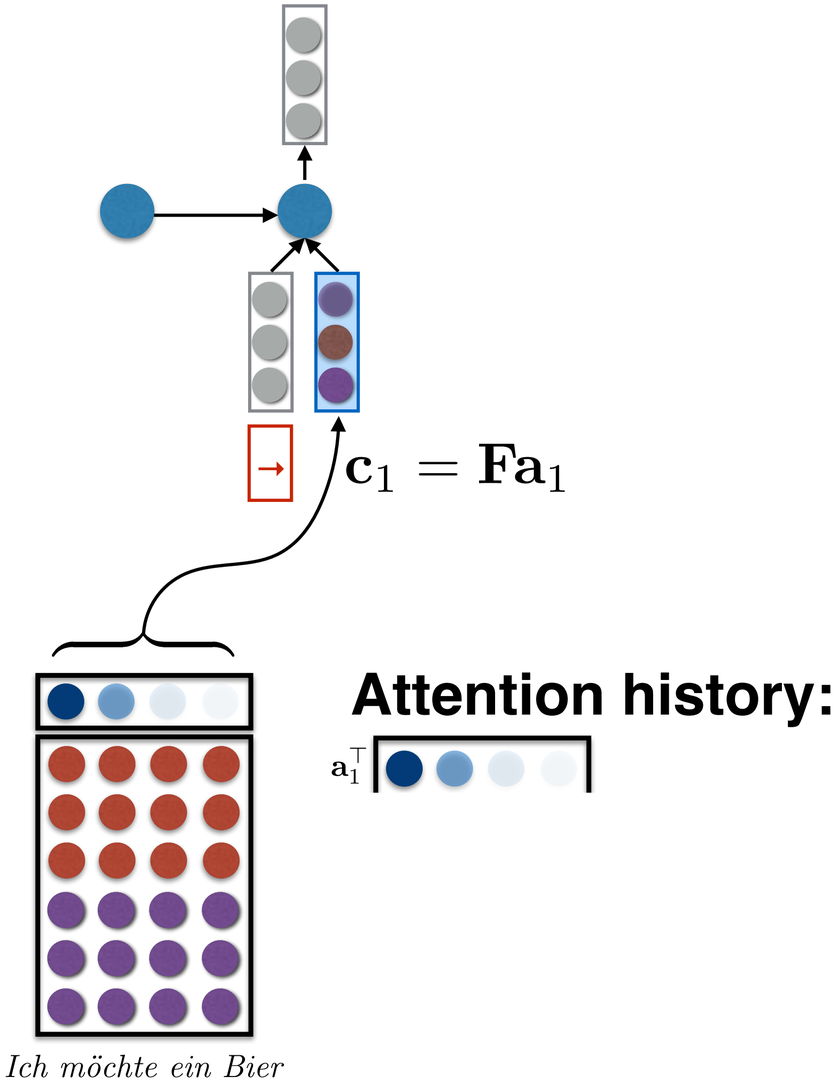


Encode S

$$\mathbf{f}_i = \mathbf{x}_i$$

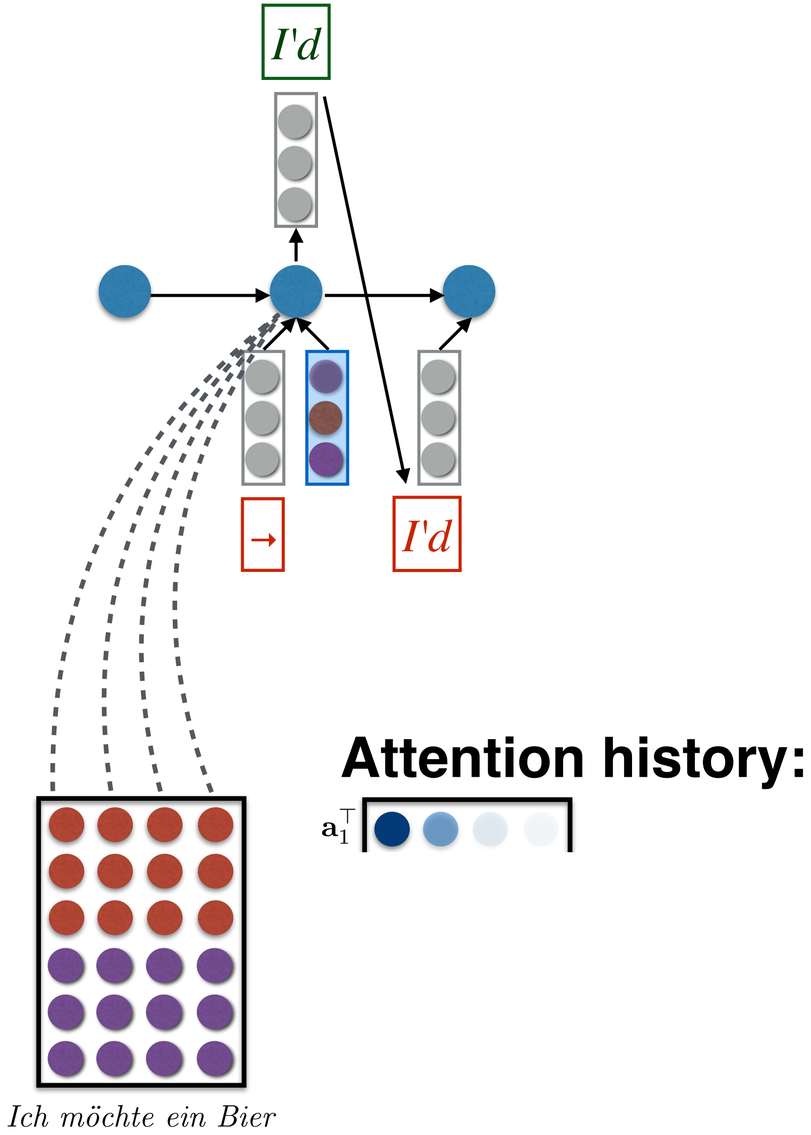






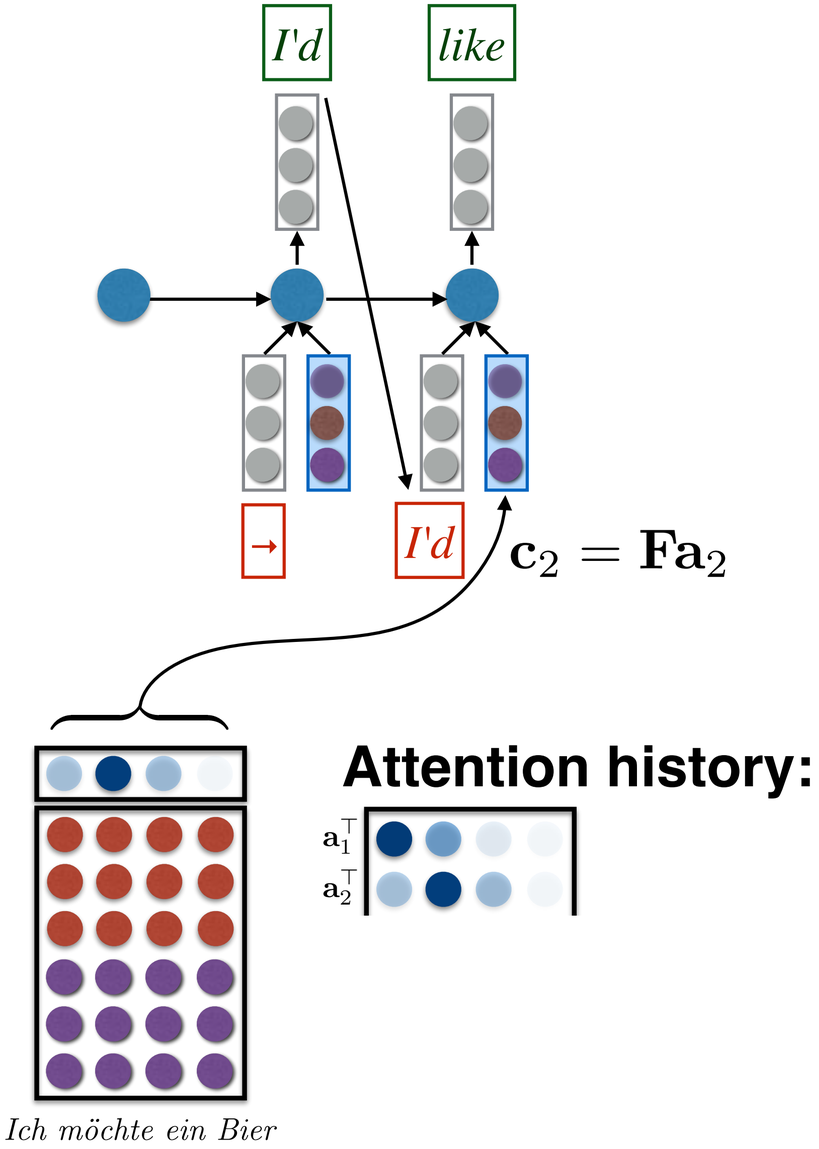
Images from Chris Dyer

Decode with Attention



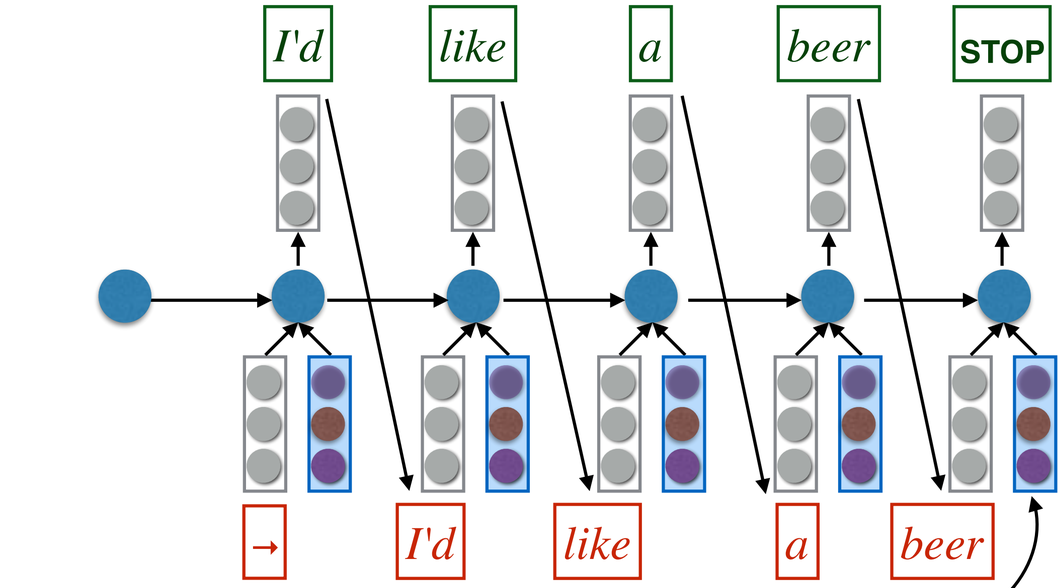
Images from Chris Dyer

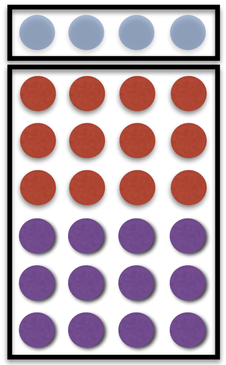
Decode with Attention



Images from Chris Dyer

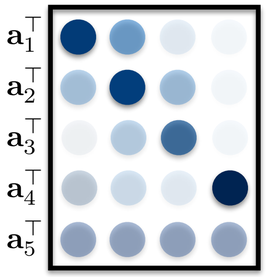
Decode with Attention





Ich möchte ein Bier

Attention history:



Images from Chris Dyer

Encoder-Decoder With Attention

- Removes bottleneck of vanilla Encoder-Decoder
- Weakens correlation between sentence length and quality
- Improves adequacy (but not completely)
- Encoder-decoder with attention became SOTA in 2015

Application Specific Features

- Length normalization
- Coverage penalty

Length Normalization

- Overcome bias towards shorter sentences
- Normalize probabilities of candidates by their length L(e)

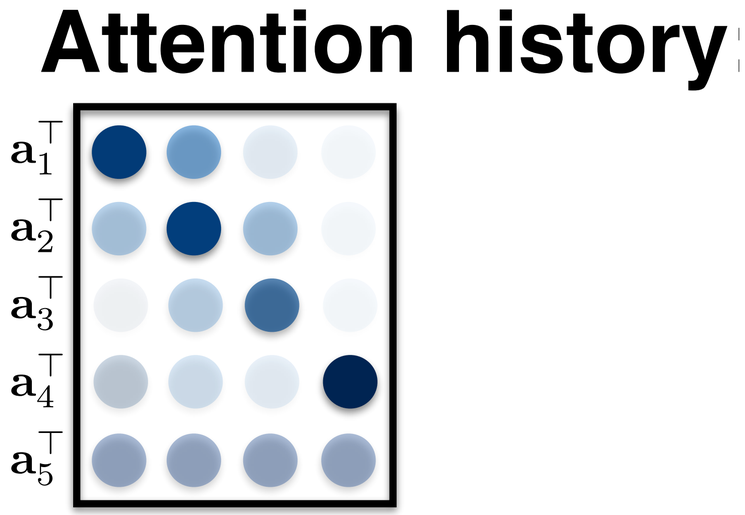
$$e^* = argmax_e \quad \frac{\Pr(e|f)}{L(e)}$$

- *Explicitly model length

$$e^* = argmax_e \Pr(L(e)|\, f)\ \Pr(e|f)$$

Coverage Penalty

- Make sure all source words were translated
- Use $\sum_j \alpha_{i,j}$ as indicator of how much we translated the *i*-th source word



$$e^* = argmax_e \quad \frac{\log \Pr(e|f)}{L(e)} + C(e,f)$$

How good is NMT?

- Significantly more fluent: lemmatized and raw BLEU scores are closer

The animal didn’t cross the street because it was too tired

Животное не пересечь улицу потому, что он слишком устал

Животное не перешло улицу потому, что оно было слишком уставшим

How good is NMT?

› Rather amazingly good at translation from very distant languages

私は交通がとても重かったので、道路を渡ることはできませんでした。

I couldn't cross the road because the traffic was so heavy.

How good is NMT?

› Rather amazingly good at translation from very distant languages

私は交通がとても重かったので、道路を渡ることはできませんでした。

I couldn't cross the road because the traffic was so heavy.

How good is NMT?

› Rather amazingly good at translation from very distant languages

私は交通がとても重かったので、道路を渡ることはできませんでした。

I couldn't cross the road because the traffic was so heavy.

How good is NMT?

› Rather amazingly good at translation from very distant languages

私は交通がとても重かったので、道路を渡ることはできませんでした。

I couldn't cross the road because the traffic was so heavy.

How good is NMT?

› Rather amazingly good at translation from very distant languages

私は**交通が**とても重かったので、道路を渡ることはできませんでした。

I couldn't cross the road because **the traffic** was so heavy.

How good is NMT?

› Rather amazingly good at translation from very distant languages

私は交通が**とても重かった**ので、道路を渡ることはできませんでした。

I couldn't cross the road because the traffic **was so heavy**.

How good is NMT?

› Rather amazingly good at translation from very distant languages

私は交通がとても重かった**ので**、道路を渡ることはできませんでした。

I couldn't cross the road **because** the traffic was so heavy.

How good is NMT?

› Rather amazingly good at translation from very distant languages

私は交通がとても重かったので、**道路を**渡ることはできませんでした。

I couldn't cross **the road** because the traffic was so heavy.

How good is NMT?

› Rather amazingly good at translation from very distant languages

私は交通がとても重かったので、道路を**渡る**ことはできませんでした。

I couldn't **cross** the road because the traffic was so heavy.

How good is NMT?

› Rather amazingly good at translation from very distant languages

私は交通がとても重かったので、道路を渡る**ことはできませんでした**。

I **couldn't** cross the road because the traffic was so heavy.

How good is NMT?

- Amazingly good at rephrasing

私は交通がとても重かったので、道路を渡ることはできませんでした。

I couldn't cross the road because the traffic was so heavy.

How good is NMT?

- Amazingly good at rephrasing

私~~は~~交通がとても重かったので、道路を渡ることはできませんでした。

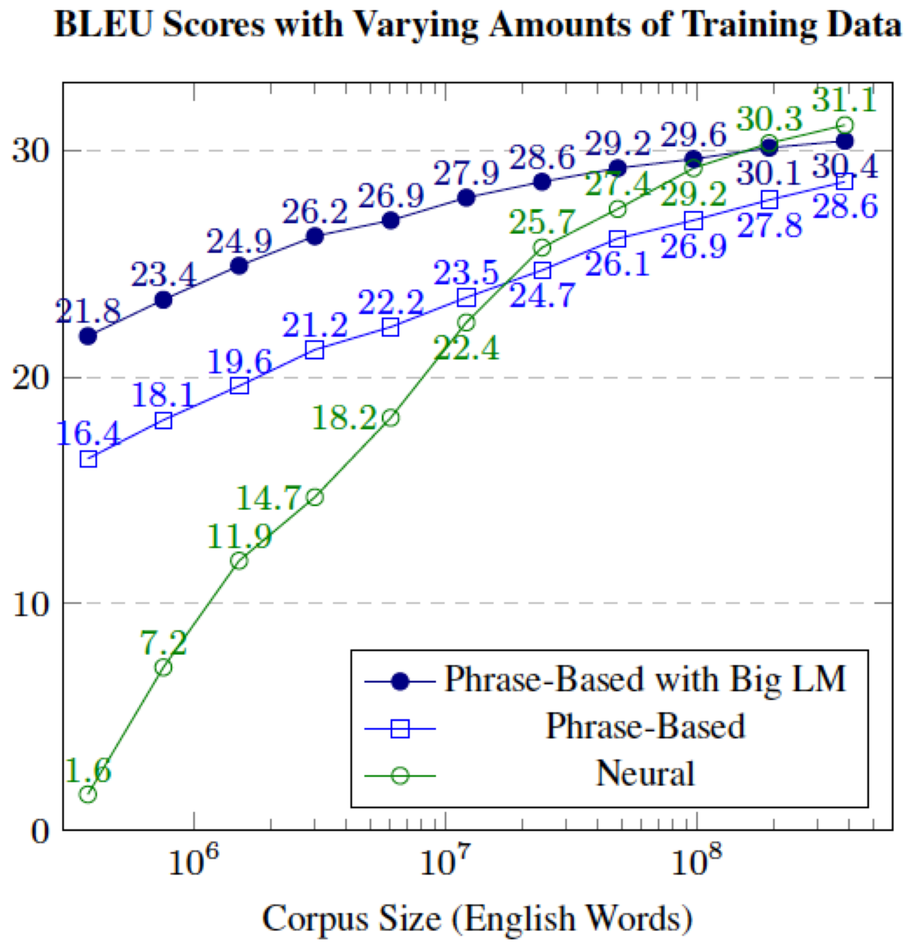
The traffic was so heavy that it was not possible to cross the road.

What Problems does NMT have?

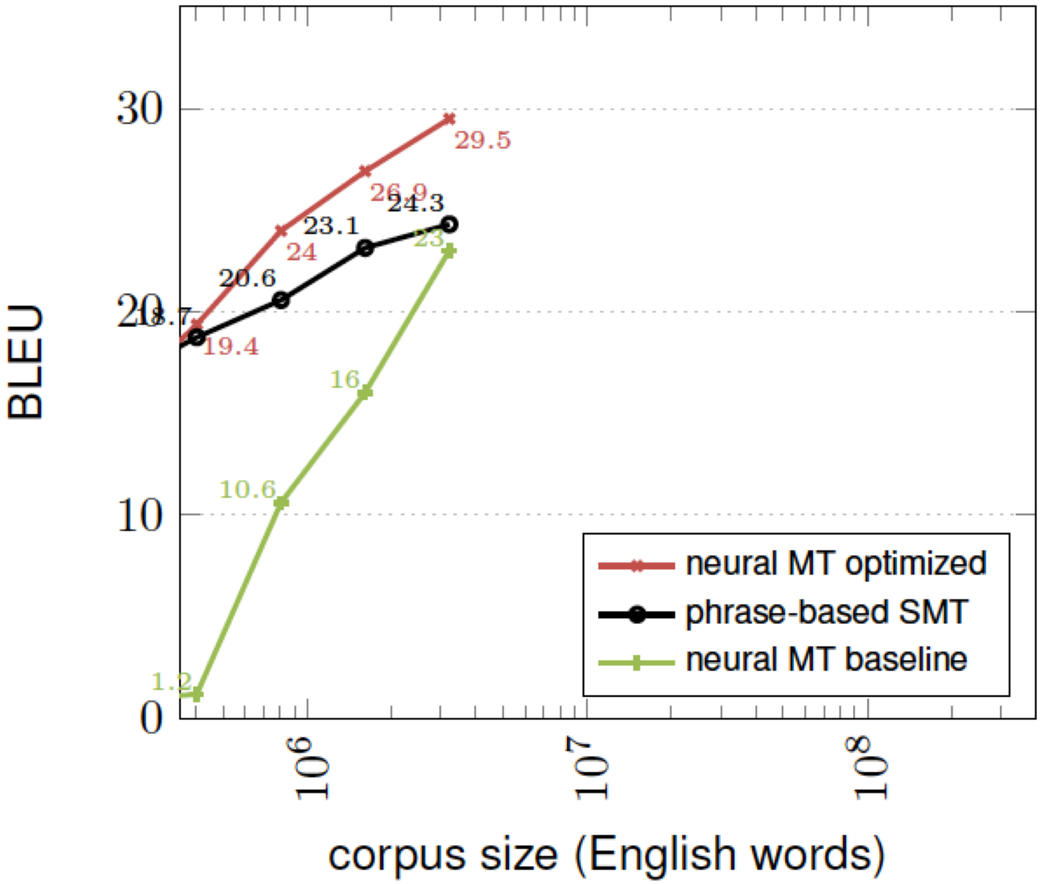
- Adequacy: word sense disambiguation
- Long sentences: no longer a noticeable issue
- Low resource setting: just needs better optimization
- Label bias?

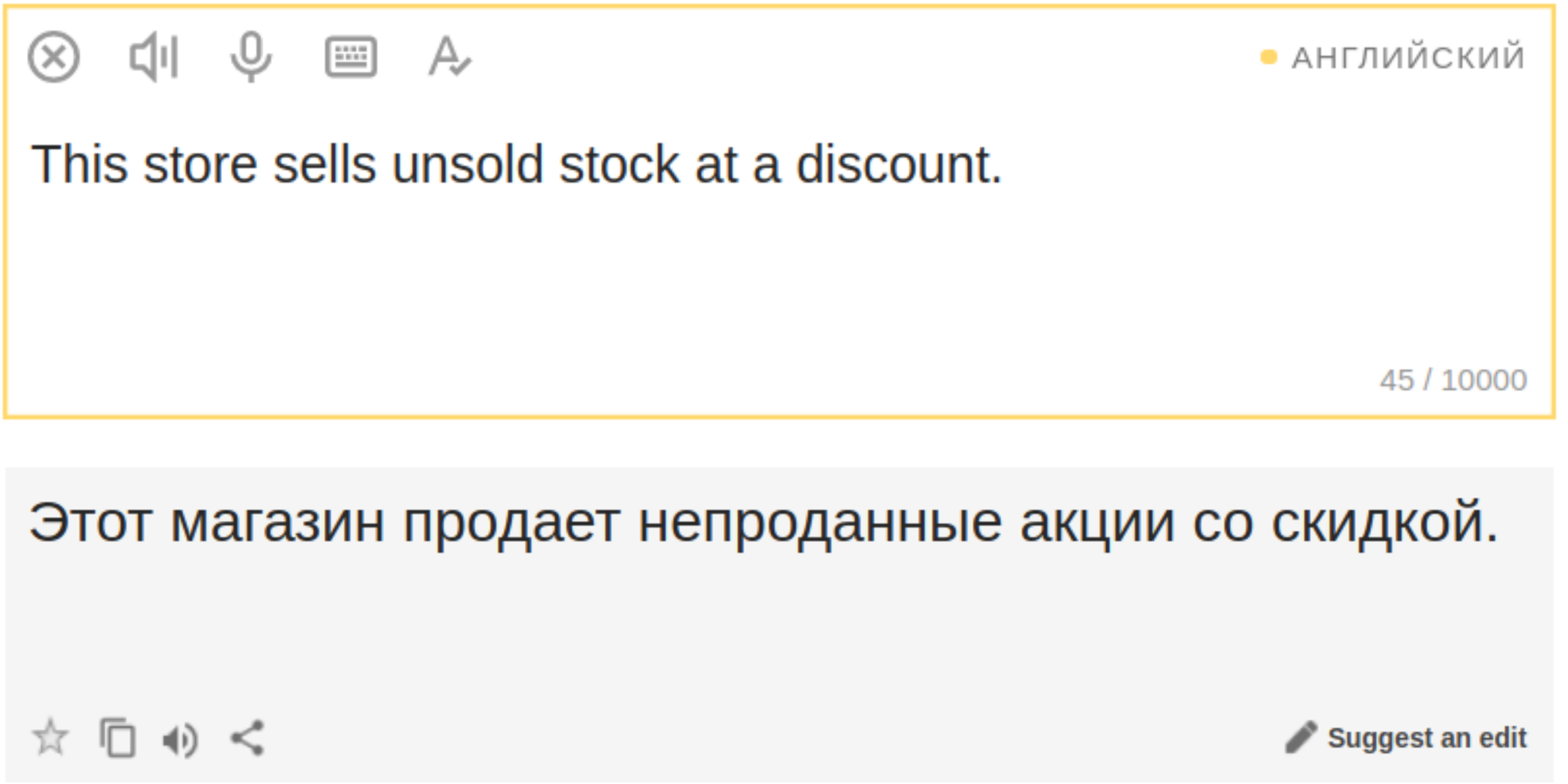
Low Resource Setting (Sennrich 2019)

[Koehn and Knowles, 2017]

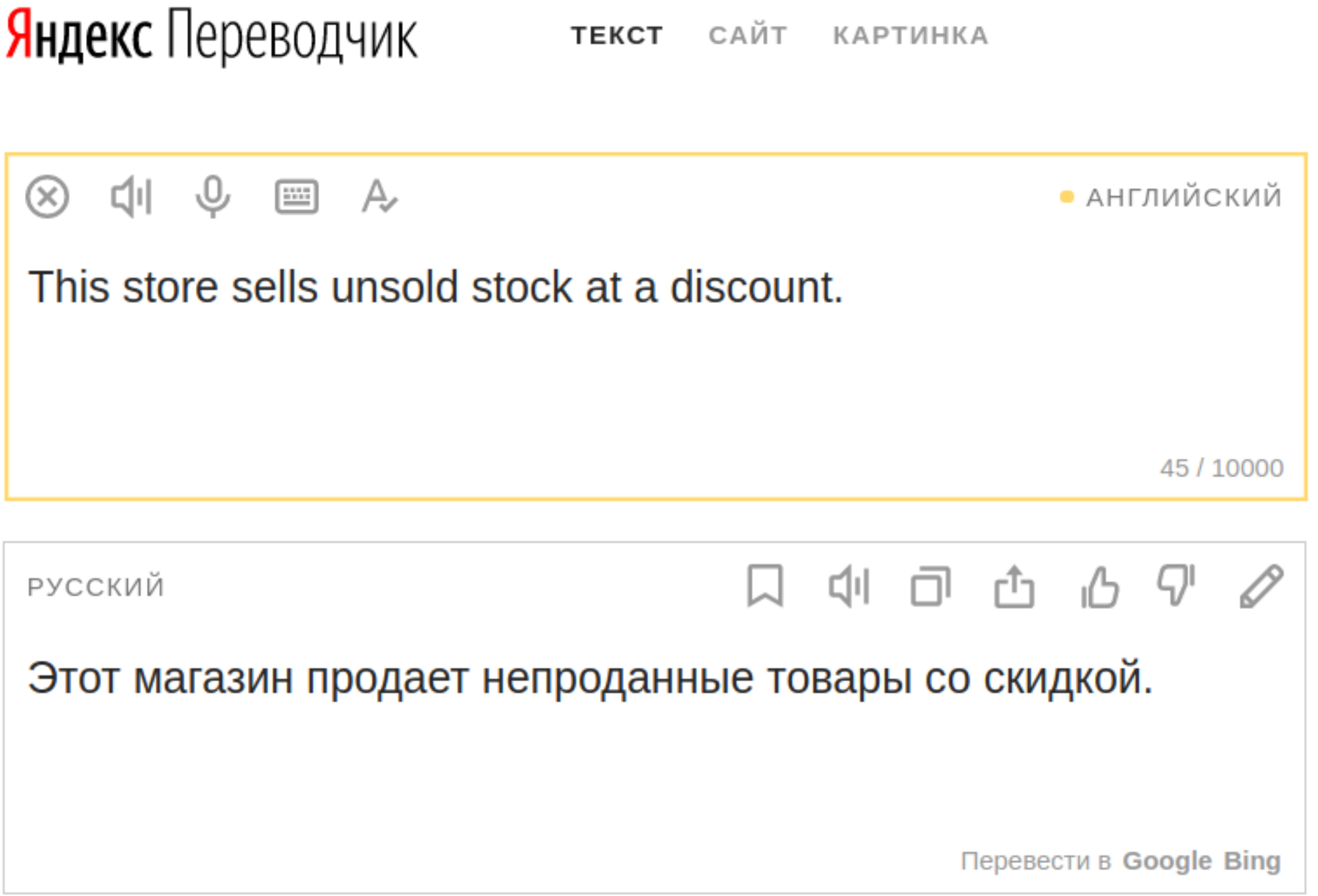


our experiments

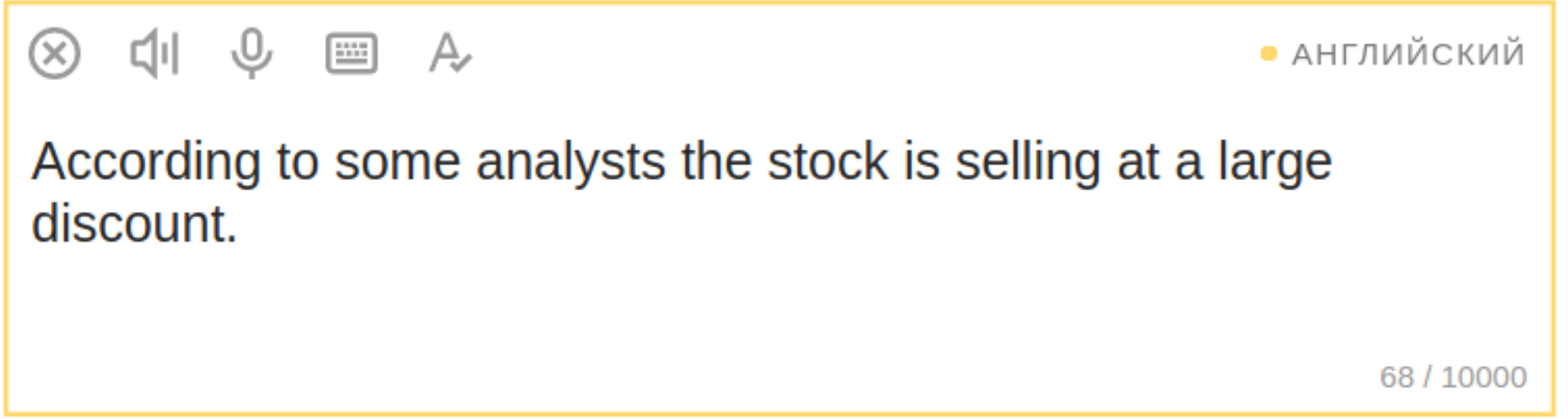



Яндекс Переводчик
ТЕКСТ
САЙТ
КАРТИНКА
АНГЛИЙСКИЙ
This store sells unsold stock at a discount.
45 / 10000
Этот магазин продает непроданные акции со скидкой.
Suggest an edit

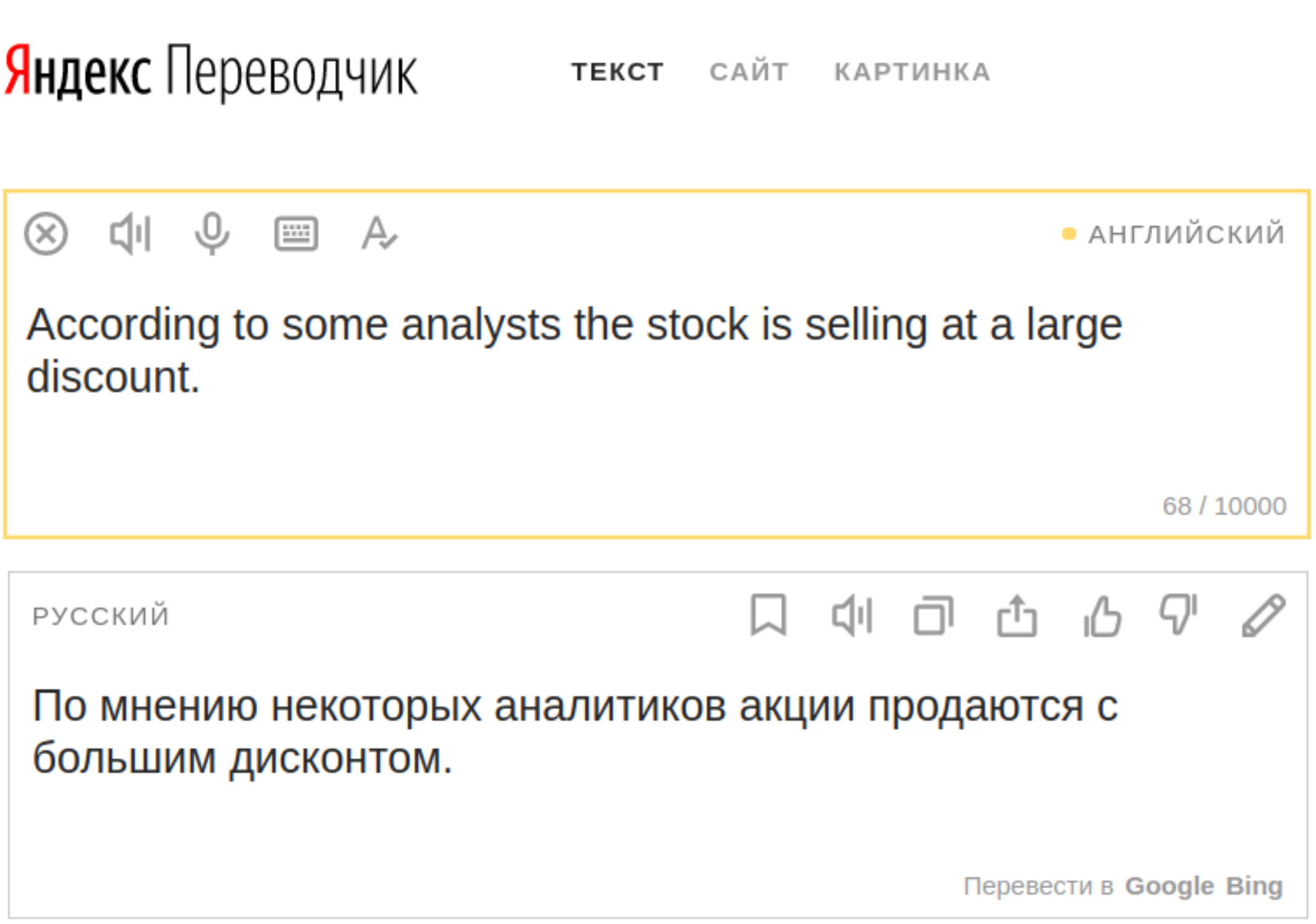
RNN



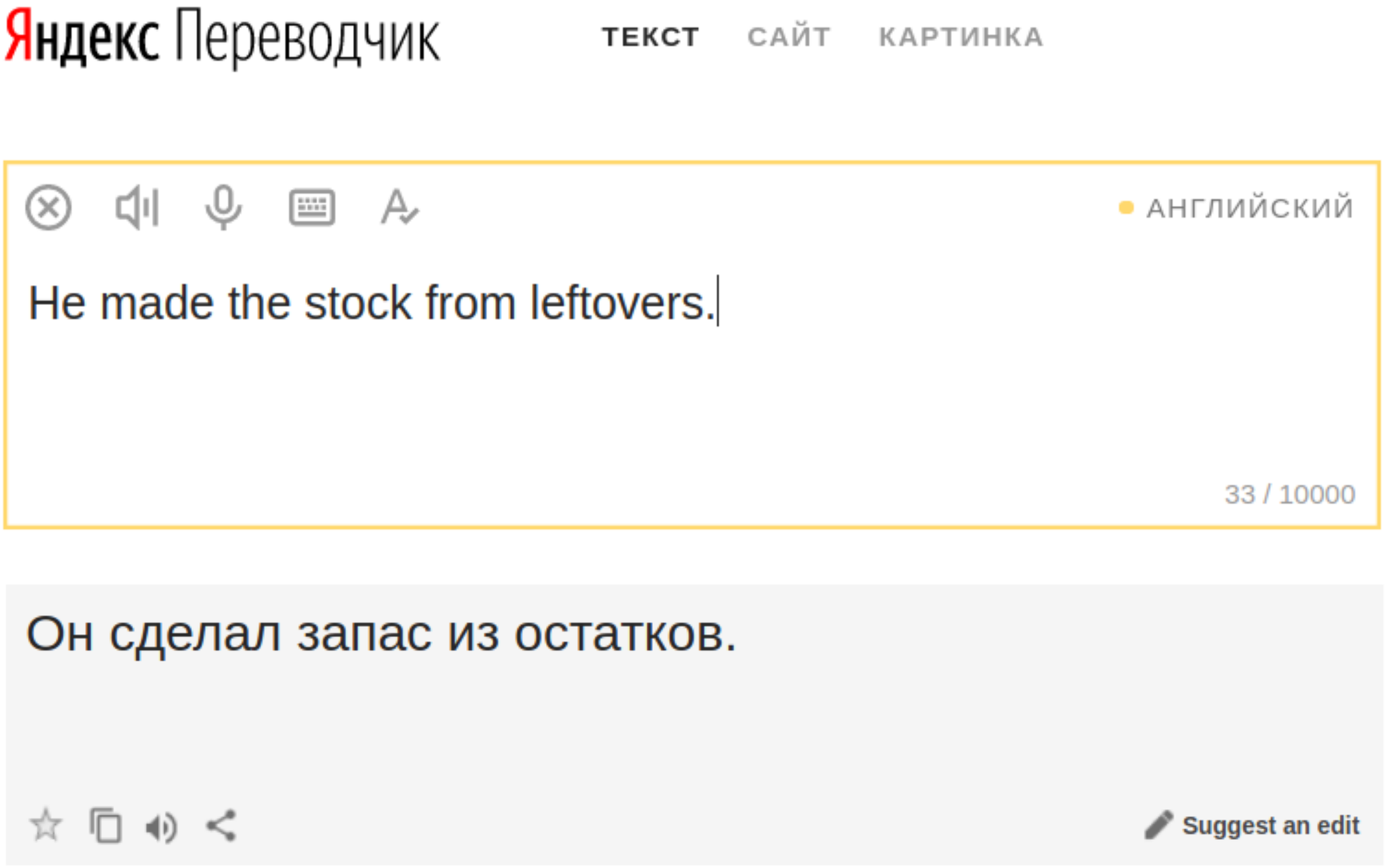
Transformer


Яндекс Переводчик
ТЕКСТ
САЙТ
КАРТИНКА
АНГЛИЙСКИЙ
According to some analysts the stock is selling at a large discount.
68 / 10000
По мнению некоторых аналитиков, акции продаются с большой скидкой.
Suggest an edit

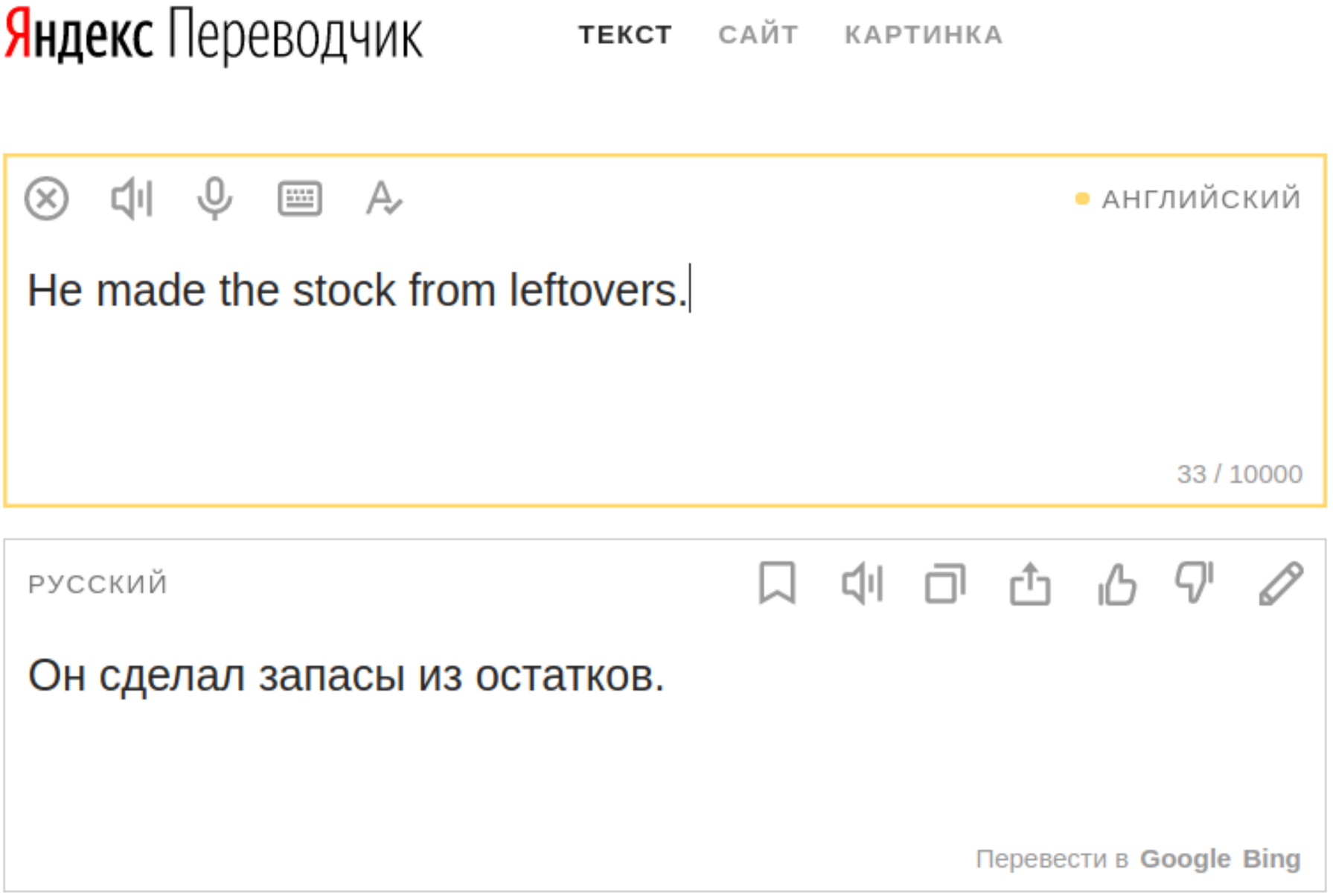
RNN



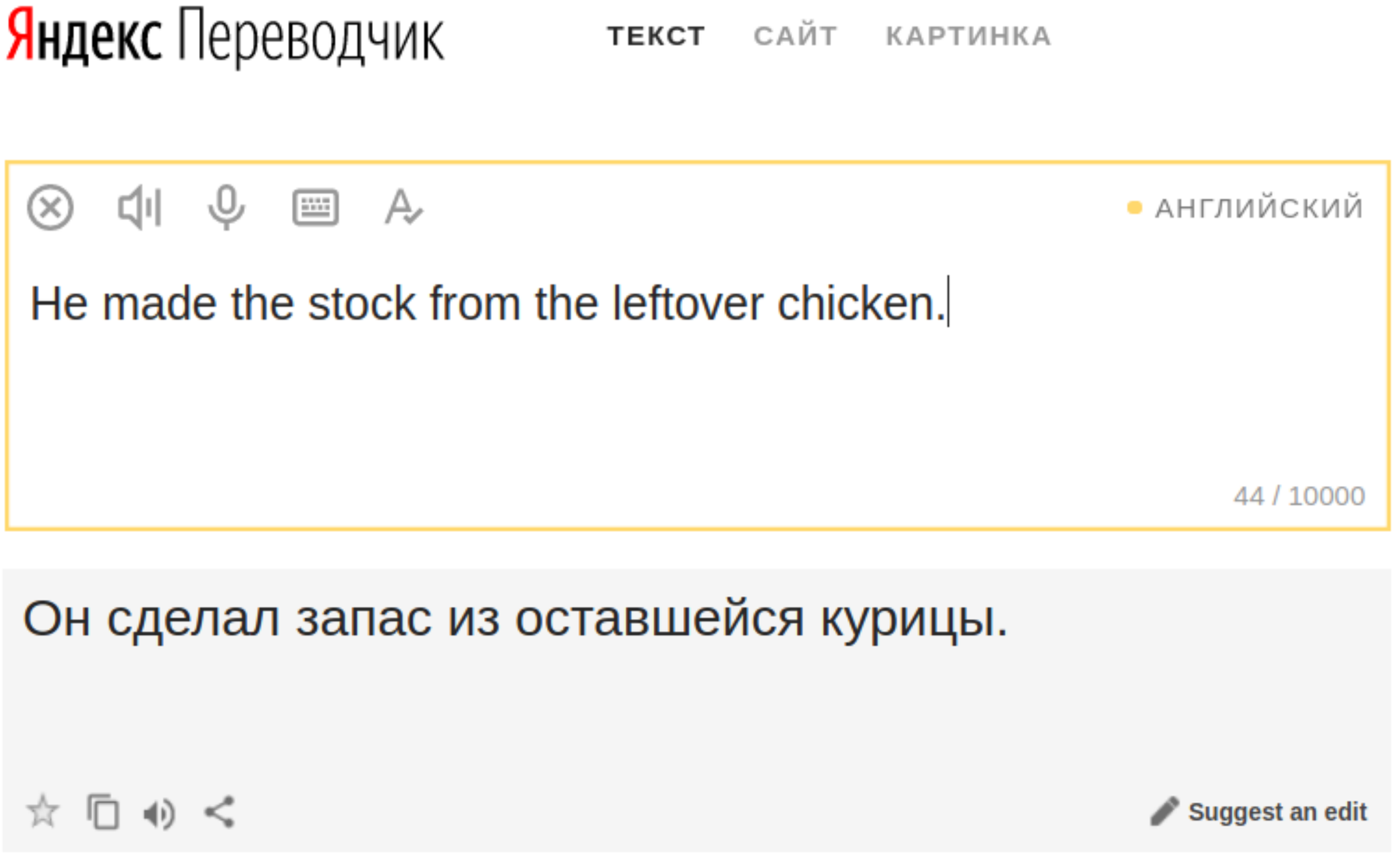
Transformer


Яндекс Переводчик
ТЕКСТ
САЙТ
КАРТИНКА
АНГЛИЙСКИЙ
He made the stock from leftovers.
33 / 10000
Он сделал запас из остатков.
Suggest an edit

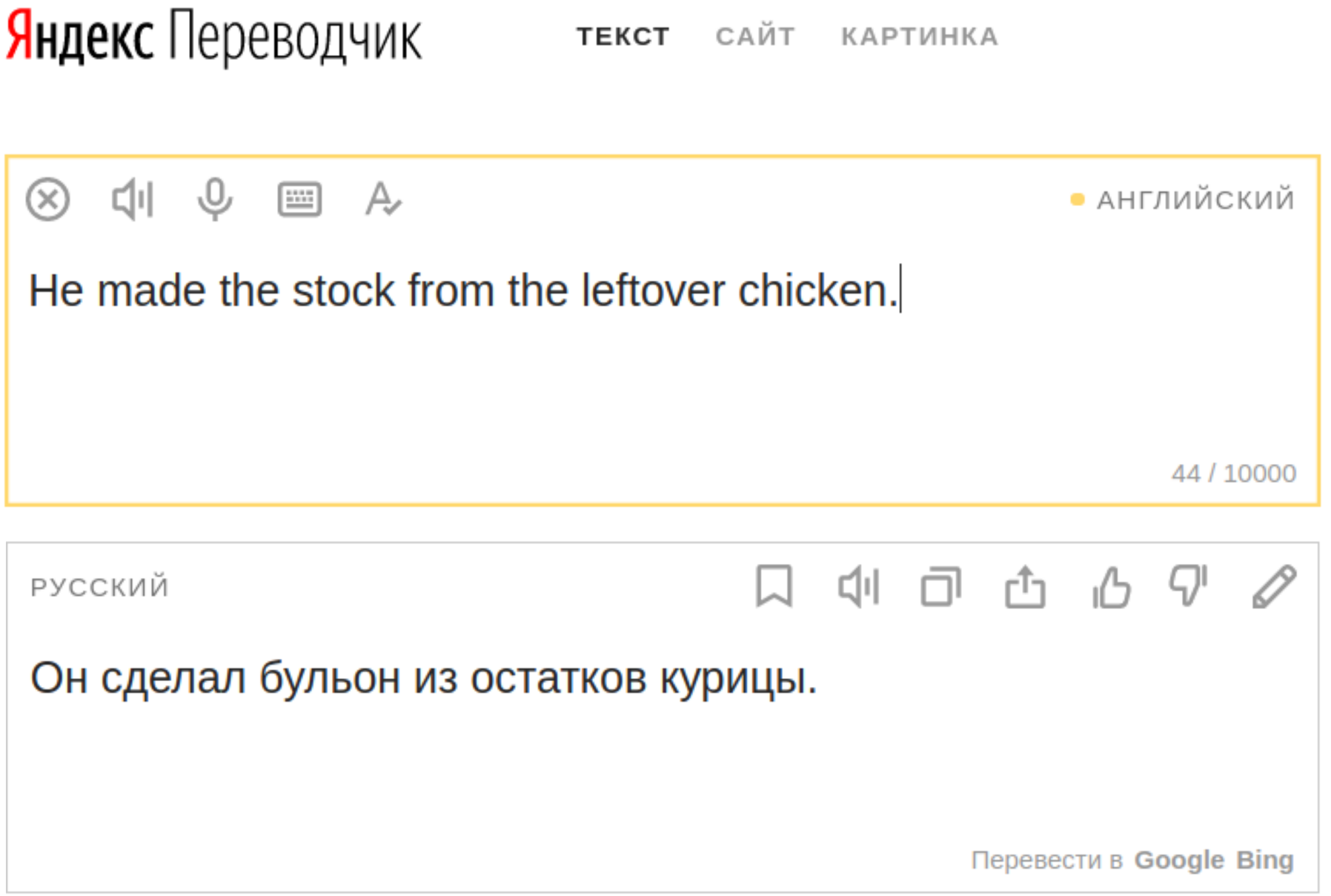
RNN



Transformer


Яндекс Переводчик
ТЕКСТ
САЙТ
КАРТИНКА
АНГЛИЙСКИЙ
He made the stock from the leftover chicken.
44 / 10000
Он сделал запас из оставшейся курицы.
Suggest an edit

RNN



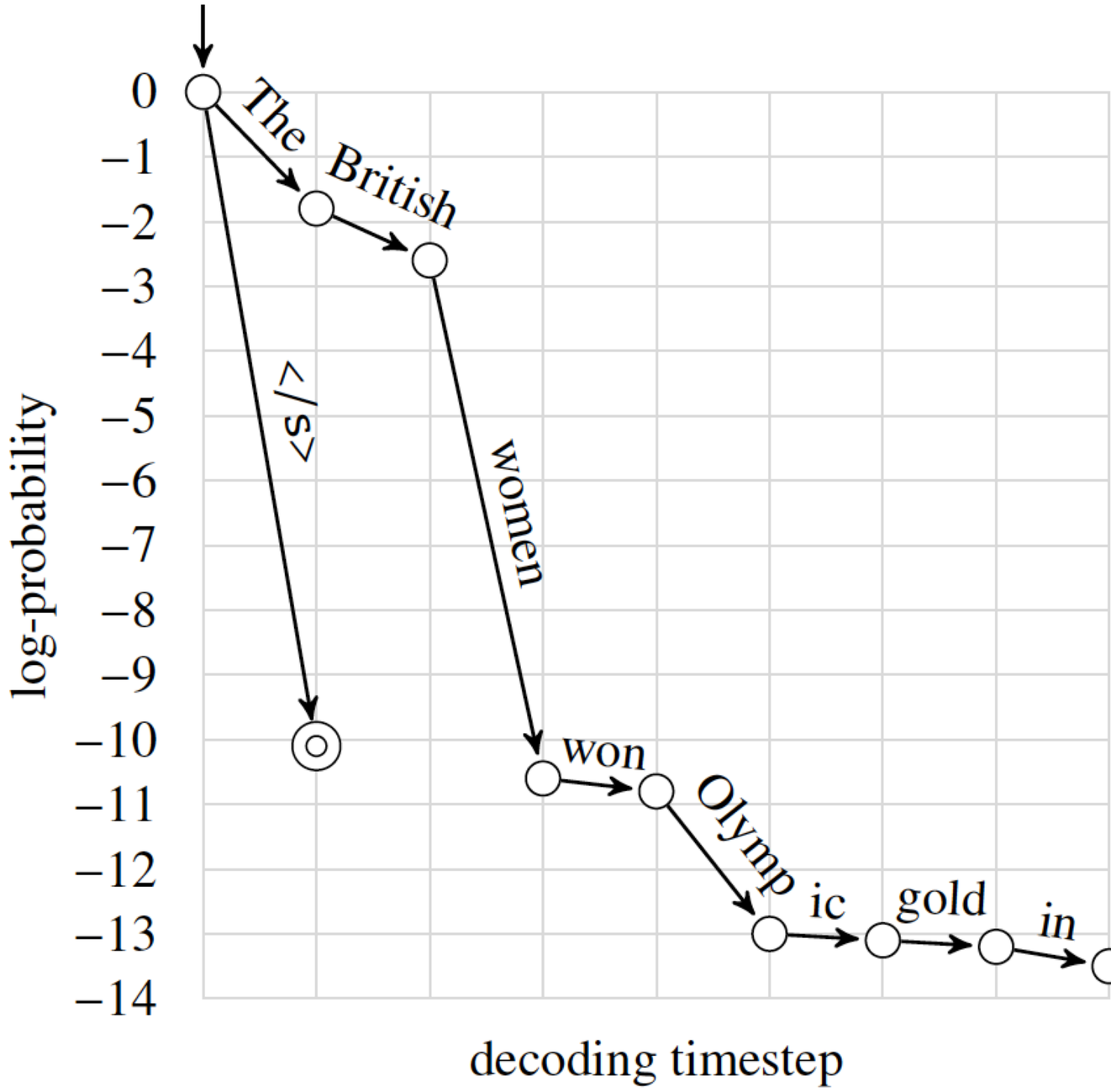
Transformer

Large Beam Hurts

- NMT has a much more powerful model
- Increasing the beam size should reduce search errors …
- Lack of diversity?
- Lack of global normalization?

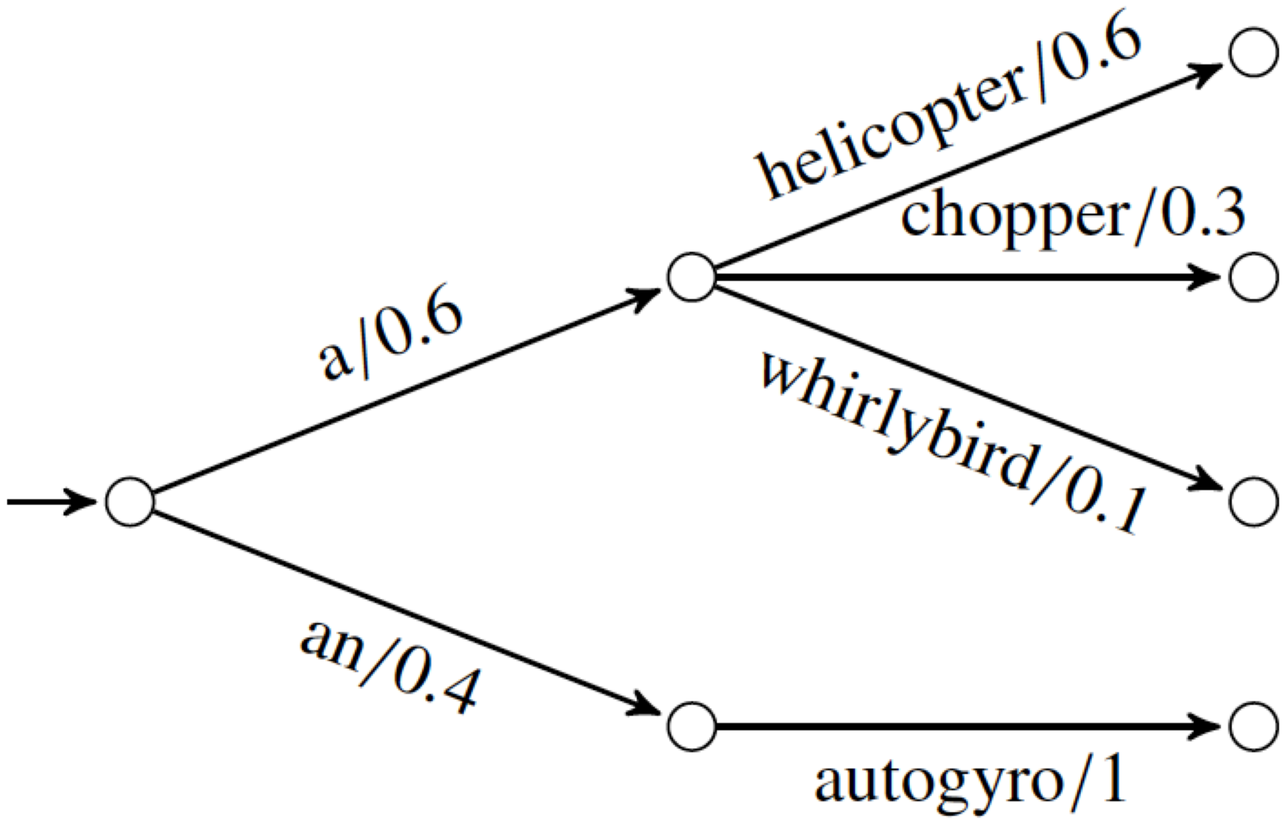
Length Bias?

Probability of hypothesis with </s> is upper-bound on probability assigned to rest of sentence

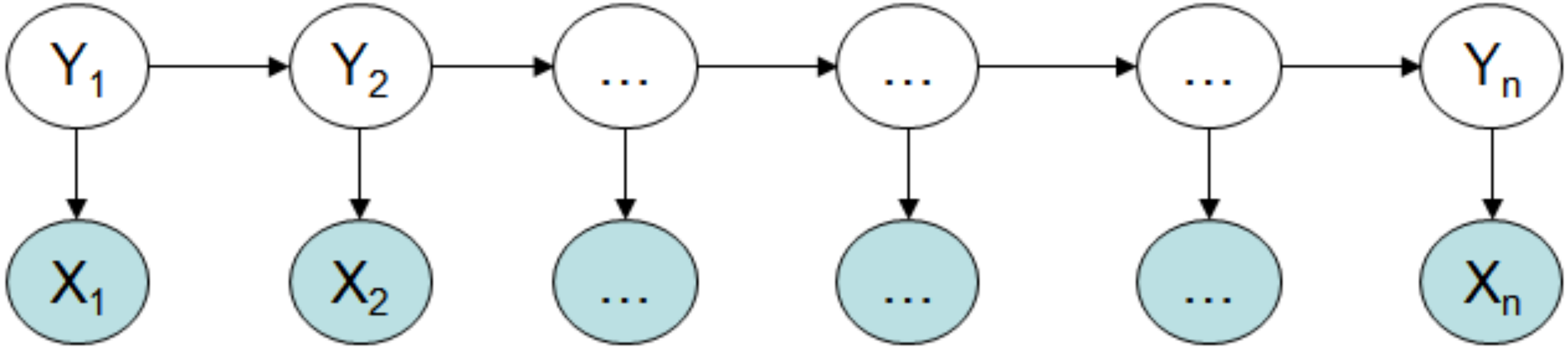


Label Bias?

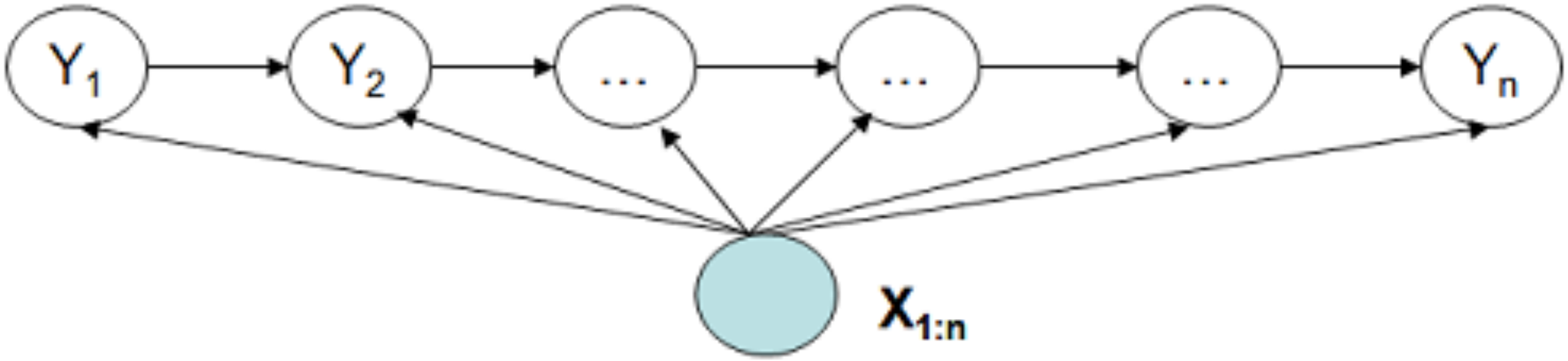
Locally normalized conditional models can ignore inputs



Label Bias?

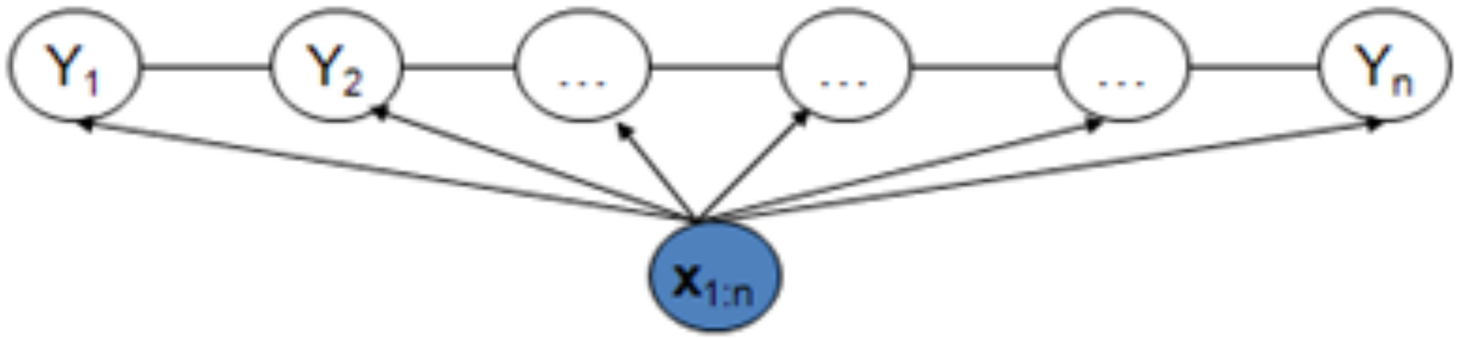


Label Bias?



$$P(\mathbf{y}_{1:n}|\mathbf{x}_{1:n}) = \prod_{i=1}^{n} P(y_i|y_{i-1},\mathbf{x}_{1:n}) = \prod_{i=1}^{n} \frac{\exp(\mathbf{w}^T\mathbf{f}(y_i,y_{i-1},\mathbf{x}_{1:n}))}{Z(y_{i-1},\mathbf{x}_{1:n})}$$

Label Bias?



- Globally normalized models?
- Reinforcement learning?
- Noisy channel?