Machine Translation

Yandex School of Data Analysis
David Talbot
October 2021

Outline

Outline

- Why Machine Translation is hard

Outline

- Why Machine Translation is hard
- The Noisy Channel approach to MT (Statistical MT)

Outline

- Why Machine Translation is hard
- The Noisy Channel approach to MT (Statistical MT)
- How Neural Machine Translation changed things

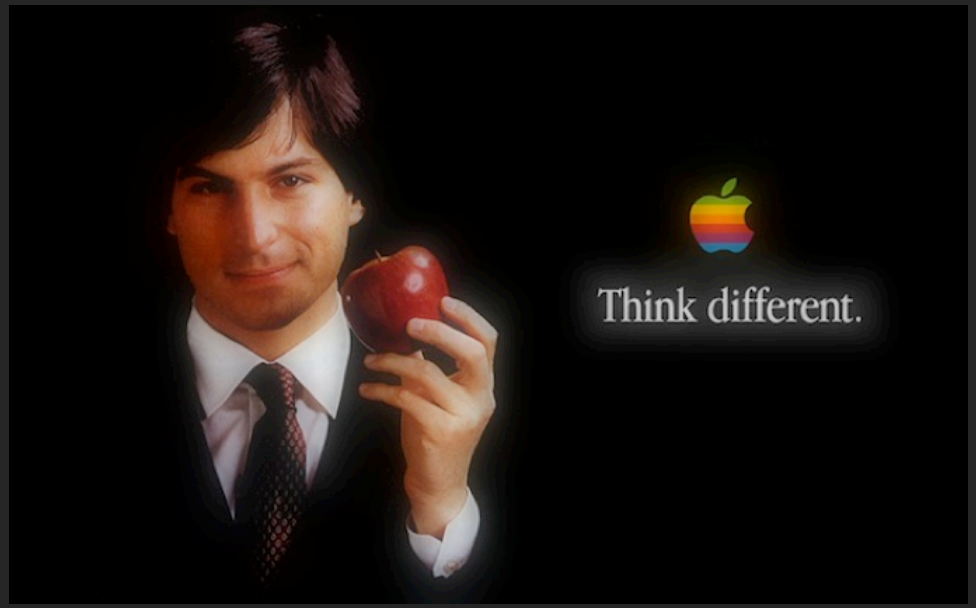
Outline

- Why Machine Translation is hard
- The Noisy Channel approach to MT (Statistical MT)
- How Neural Machine Translation changed things
- Current challenges in NMT

Why Machine Translation is Hard

Ambiguity

“A computer that understands you like your mother”



Ambiguity

“A computer that understands you like your mother”

- Компьютер, который понимает тебя, как твою мать

Ambiguity

“A computer that understands you like your mother”

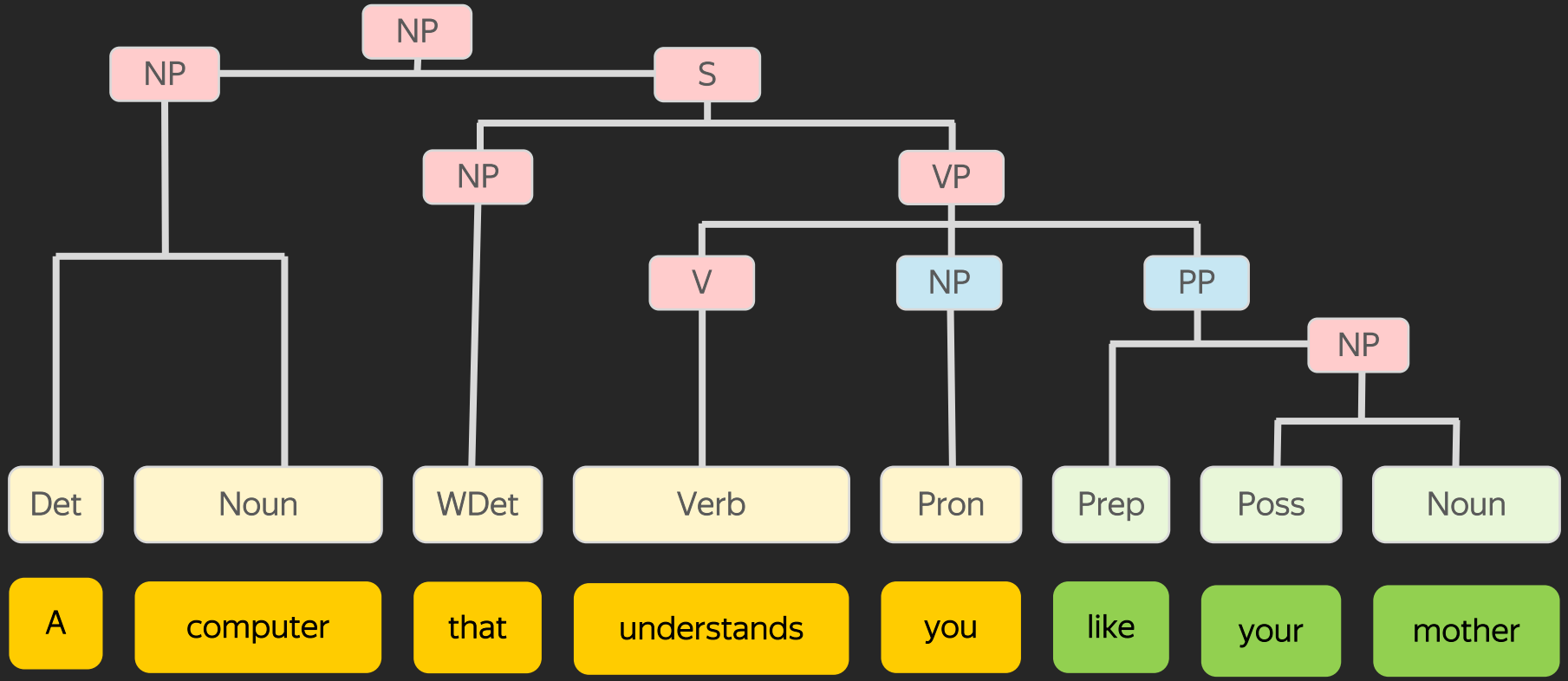
- Компьютер, который понимает тебя, как твою мать
- Компьютер, который понимает, что ты любишь твою мать

Ambiguity

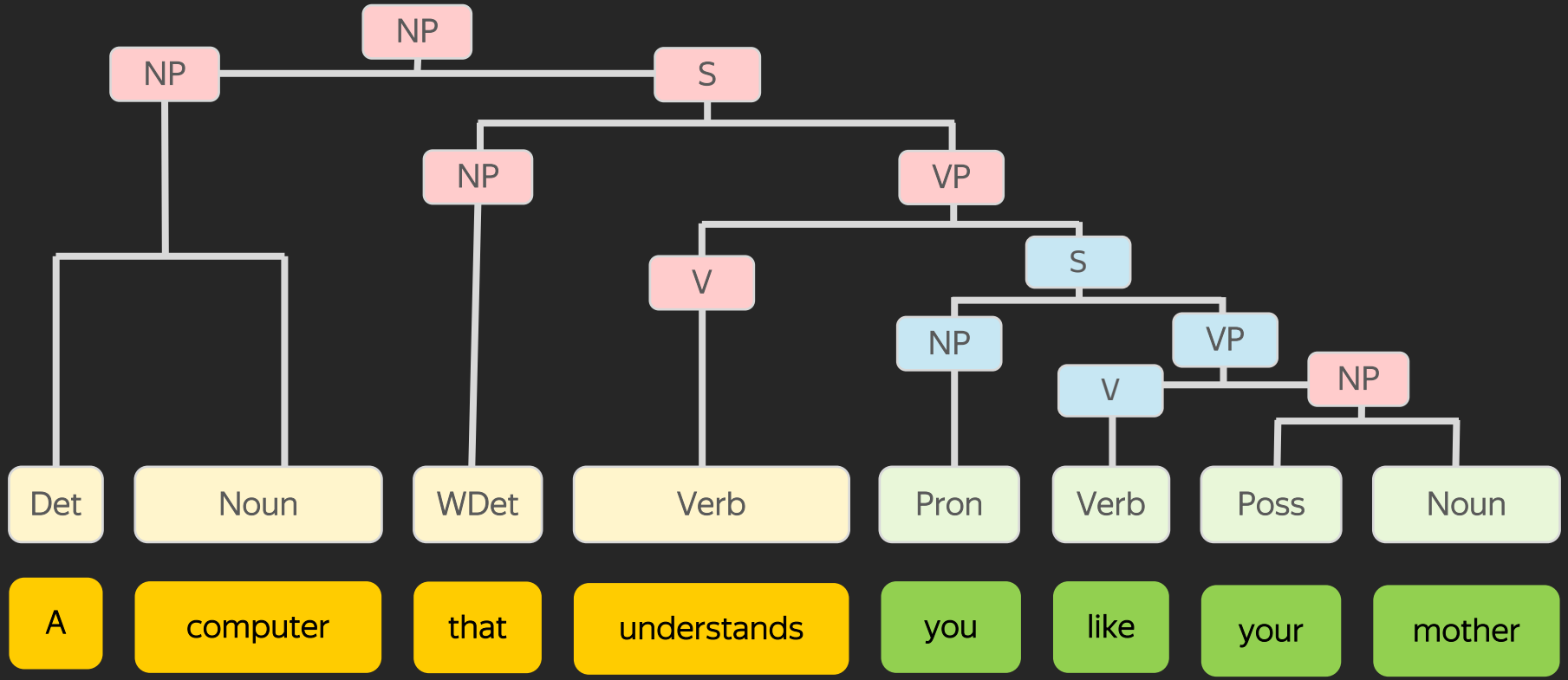
“A computer that understands you like your mother”

- Компьютер, который понимает тебя, как твою мать
- Компьютер, который понимает, что ты любишь твою мать
- Компьютер, который понимает тебя так же, как он понимает твою мать (?)

Syntactic Ambiguity



Syntactic Ambiguity



AI Completeness

What's the hardest word to translate in these sentences?

"The animal didn't cross the road because it was too wide."

AI Completeness

What's the hardest word to translate in these sentences?

"The animal didn't cross the road because it was too wide."
"The animal didn't cross the road because it was too tired."

AI Completeness

What's the hardest word to translate in these sentences?

"The animal didn't cross the road because it was too wide."
"The animal didn't cross the road because it was too tired."

AI Completeness

What's the hardest word to translate in these sentences?

"The animal didn't cross the road because it was too wide."
"The animal didn't cross the road because it was too tired."

AI Completeness

What's the hardest word to translate in these sentences?

"The animal didn't cross the road because it was too wide."
"The animal didn't cross the road because it was too tired."

AI Completeness

What's the hardest word to translate in these sentences?

"The animal didn't cross the road because it was too wide."
"The animal didn't cross the road because it was too tired."

AI Completeness

What's the hardest word to translate in these sentences?

"The animal didn't cross the road because it was too wide."
"The animal didn't cross the road because it was too tired."

AI Completeness

What's the hardest word to translate in these sentences?

"The animal didn't cross the road because it was too wide."
"The animal didn't cross the road because it was too tired."

Languages Are Different

Languages Are Different

- Syntactic structure, word order, head-directionality

Languages Are Different

- Syntactic structure, word order, head-directionality
- Isolating, analytic, synthetic, fusional, etc.

Languages Are Different

- Syntactic structure, word order, head-directionality
- Isolating, analytic, synthetic, fusional, etc.
- Morphosyntactic alignment

Head-final Japanese

A computer that understands you like your mother

Head-final Japanese

A computer that understands you like your mother

お母さん　のように　理解してくれる　コンピュータ

mother [polite]	similar	understanding giving [inside group]	computer

Head-final Japanese

A computer that understands you like your mother

お母さん のように 理解してくれる コンピュータ

mother [polite]	similar	understanding giving [inside group]	computer

Head-final Japanese

A computer that understands you like your mother

お母さん　のように　理解してくれる　コンピュータ

mother [polite]	similar	understanding giving [inside group]	computer

Head-final Japanese

A computer that understands you like your mother

お母さん のように 理解してくれる コンピュータ

mother [polite]	similar	understanding giving [inside group]	computer

Head-final Japanese

A computer that understands you like your mother

お母さん　のように　理解してくれる　コンピュータ

mother [polite]	similar	understanding giving [inside group]	computer

Ergative languages

Ŋuma banaganyu.	Father returned.
Yabu ŋumaŋgu buṛan.	Father saw mother.
Ŋuma yabuŋgu buṛan.	Mother saw father.

Ergative languages

Ŋuma banaganyu.	Father returned.
Yabu ŋumaŋgu buṛan.	Father saw mother.
Ŋuma yabuŋgu buṛan.	Mother saw father.

Translate:

Ŋuma banaganyu, yabuŋgu buṛan.

Ergative languages

Ŋuma banaganyu.	Father returned.
Yabu ŋumaŋgu buṛan.	Father saw mother.
Ŋuma yabuŋgu buṛan.	Mother saw father.

Translate:

Ŋuma banaganyu, yabuŋgu buṛan.

Father returned and [he] saw mother.

Ergative languages

Ŋuma banaganyu.	Father returned.
Yabu ŋumaŋgu buṛan.	Father saw mother.
Ŋuma yabuŋgu buṛan.	Mother saw father.

Translate:

Ŋuma banaganyu, yabuŋgu buṛan.

Father returned and [he] saw mother.

Ergative languages

Ŋuma banaganyu.	Father returned.
Yabu ŋumaŋgu buṛan.	Father saw mother.
Ŋuma yabuŋgu buṛan.	Mother saw father.

Translate:

Ŋuma banaganyu, yabuŋgu buṛan.

~~Father returned and [he] saw mother.~~

Father returned and mother saw [him].

The Story So Far

The Story So Far

- Information Theory (1940s)

The Story So Far

- Information Theory (1940s)
- Rule-based Machine Translation (1950s)

The Story So Far

- Information Theory (1940s)
- Rule-based Machine Translation (1950s)
- Statistical Machine Translation (1990s)

The Story So Far

- Information Theory (1940s)
- Rule-based Machine Translation (1950s)
- Statistical Machine Translation (1990s)
- Phrase-based Machine Translation (2000s)

The Story So Far

- Information Theory (1940s)
- Rule-based Machine Translation (1950s)
- Statistical Machine Translation (1990s)
- Phrase-based Machine Translation (2000s)
- Tree-based Machine Translation (2000s)

The Story So Far

- Information Theory (1940s)
- Rule-based Machine Translation (1950s)
- Statistical Machine Translation (1990s)
- Phrase-based Machine Translation (2000s)
- Tree-based Machine Translation (2000s)
- Neural Machine Translation (2010s)

Progress in MT

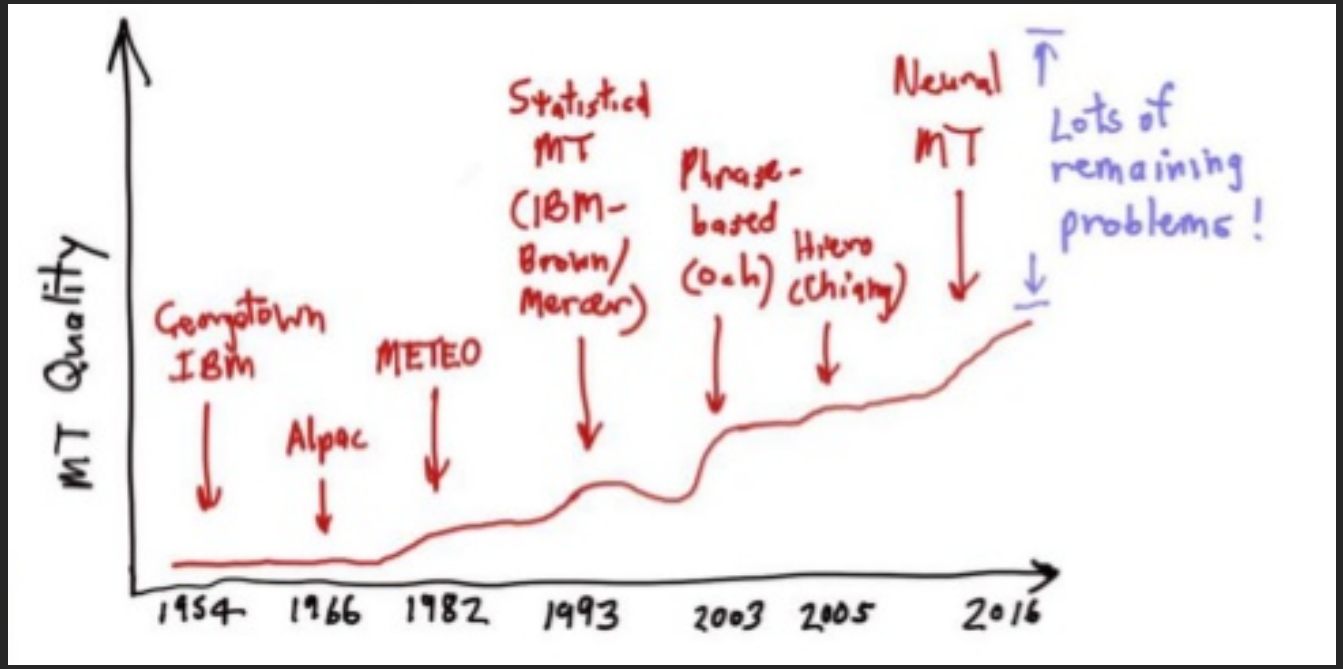
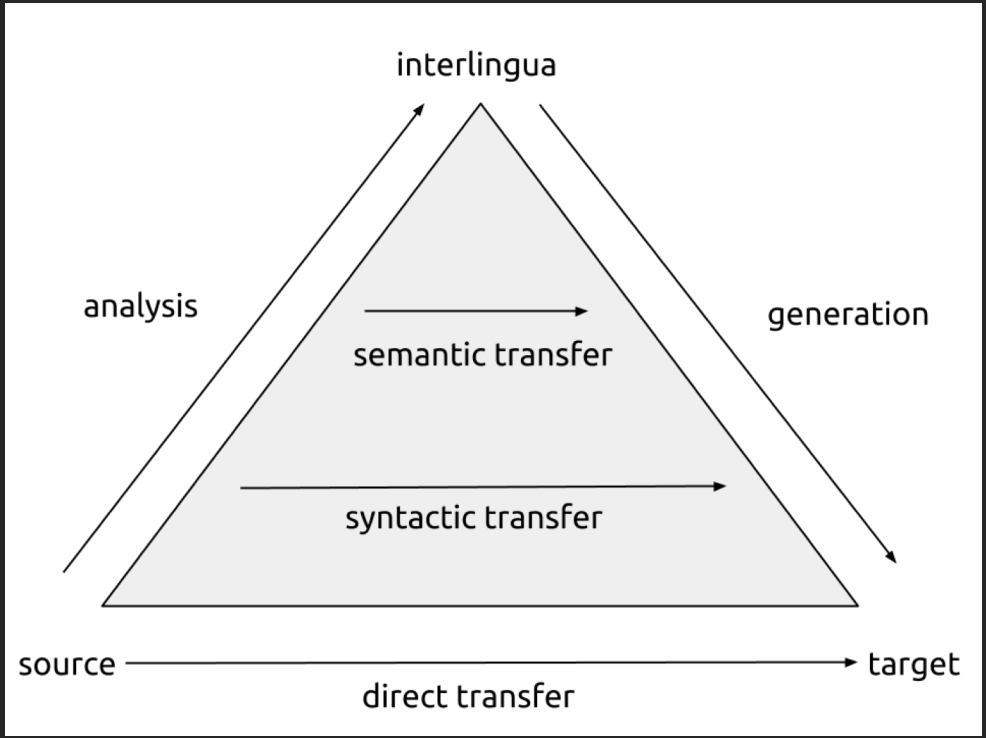
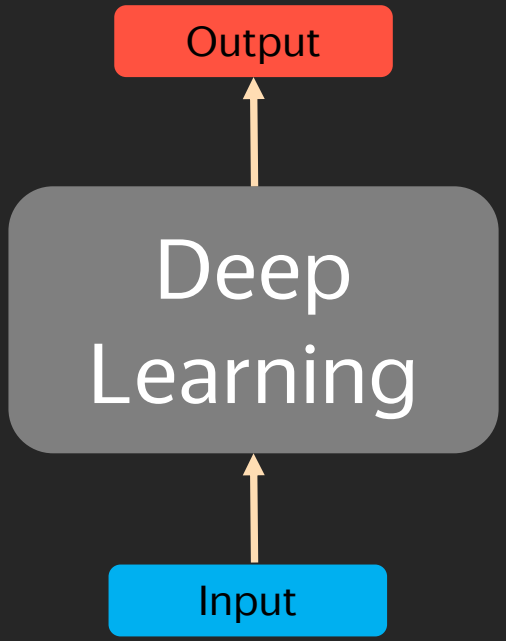


Image by Chris Manning

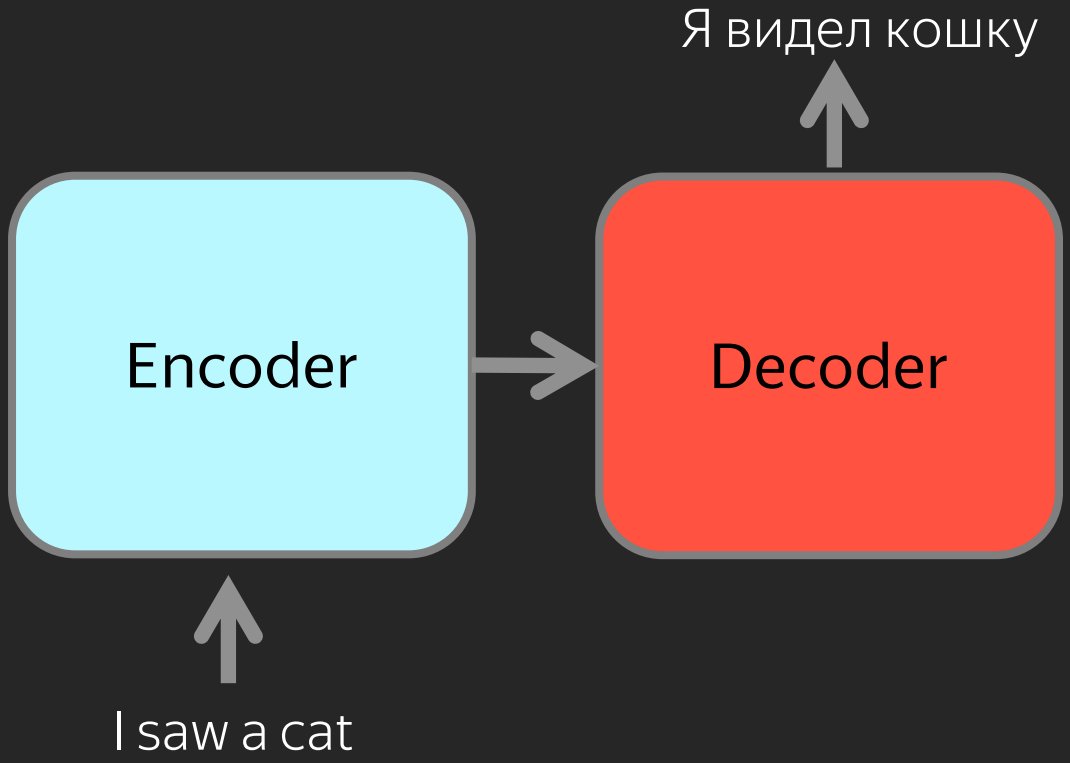
Vauquois Pyramid



Neural Machine Translation



Neural Machine Translation



The Noisy Channel Approach to MT

The Mathematics of Machine Translation (Brown 1993)

The Mathematics of Machine Translation (Brown 1993)

First NLP paper to propose:

The Mathematics of Machine Translation (Brown 1993)

First NLP paper to propose:

1. To model translation probabilistically

The Mathematics of Machine Translation (Brown 1993)

First NLP paper to propose:

1. To model translation probabilistically

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

where f is an observed French sentence and e is a translation to English.

The Mathematics of Machine Translation (Brown 1993)

First NLP paper to propose:

1. To model translation probabilistically
2. To decompose the model into a ‘source’ and ‘channel’

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

where f is an observed French sentence and e is a translation to English.

The Mathematics of Machine Translation (Brown 1993)

First NLP paper to propose:

1. To model translation probabilistically
2. To decompose the model into a ‘source’ and ‘channel’

$$e^* = argmax_e \; Pr(e|f)$$
$$= argmax_e \; \underbrace{Pr(e)}_{\text{source}} \; \underbrace{Pr(f|e)}_{\text{channel}}$$

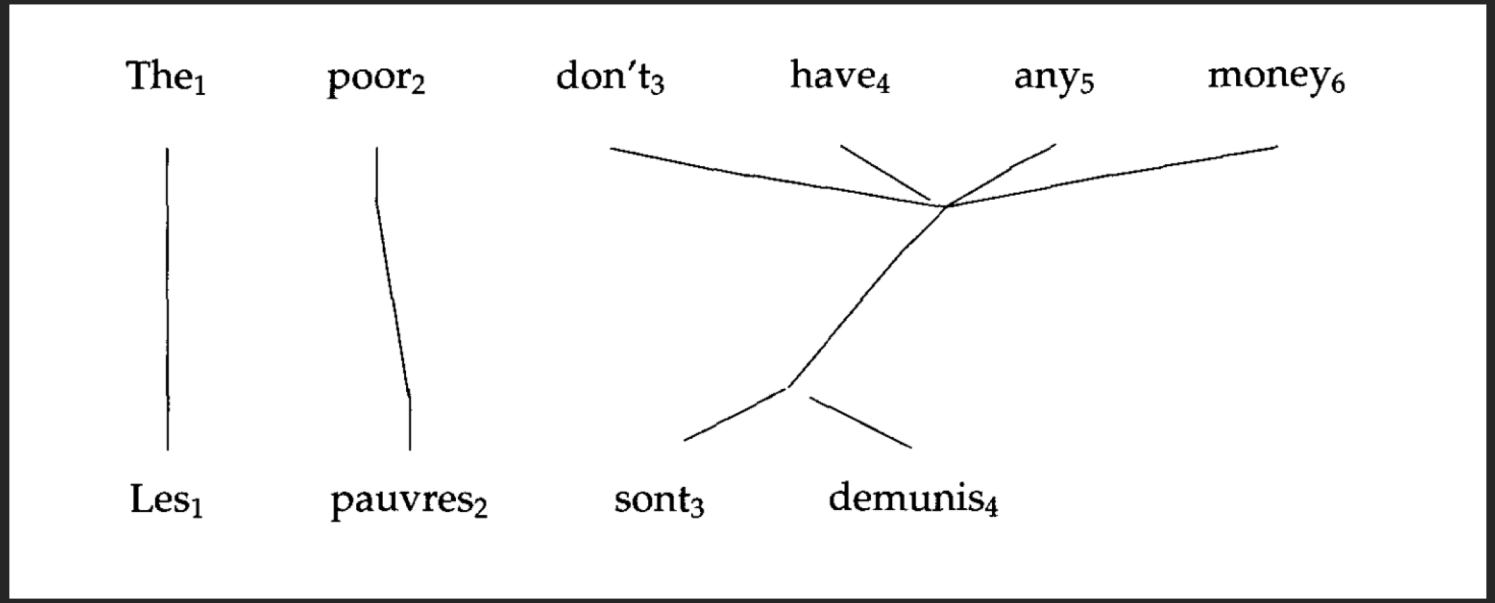
The Mathematics of Machine Translation (Brown 1993)

First NLP paper to propose:

1. To model translation probabilistically
2. To decompose the model into a ‘source’ and ‘channel’
3. To estimate the ‘channel’ model from bilingual parallel corpora

$$e^* = argmax_e \; Pr(e|f)$$
$$= argmax_e \; \underbrace{Pr(e)}_{\text{source}} \; \underbrace{Pr(f|e)}_{\text{channel}}$$

The Mathematics of Machine Translation (Brown 1993)



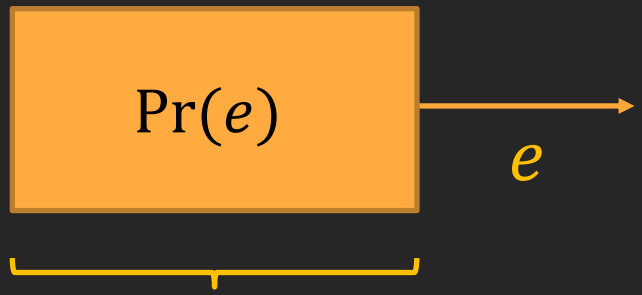
The Noisy Channel

The Noisy Channel

f

observation

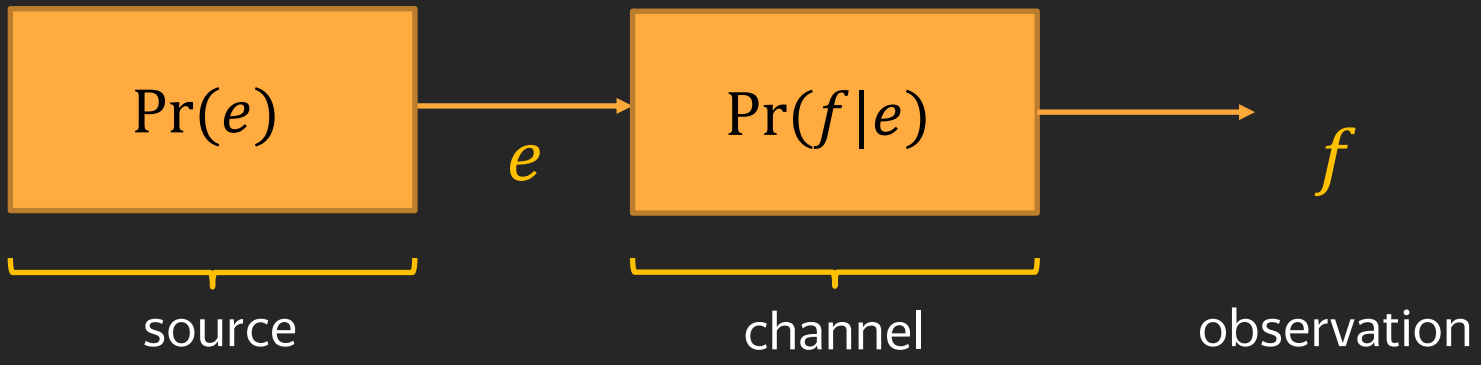
The Noisy Channel



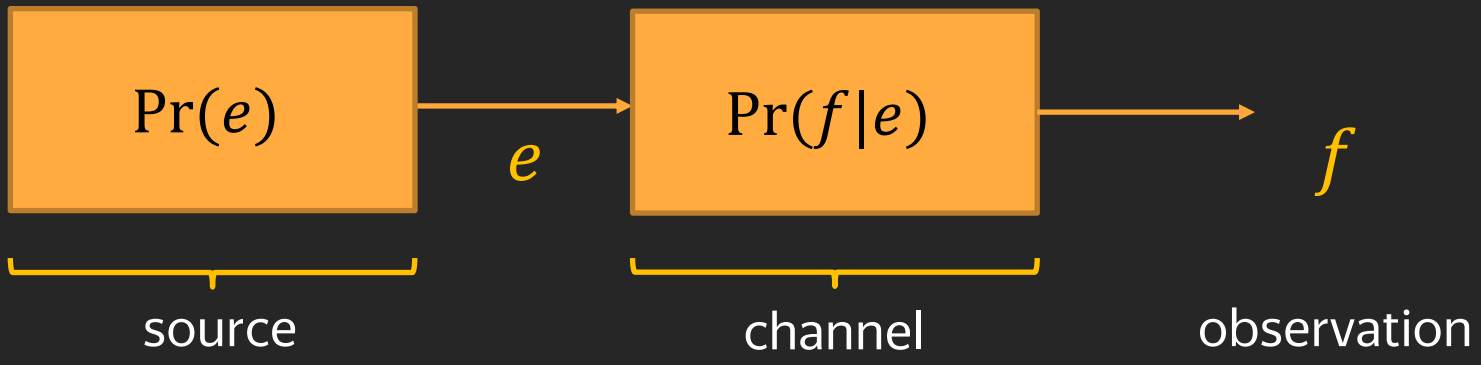
f

observation

The Noisy Channel

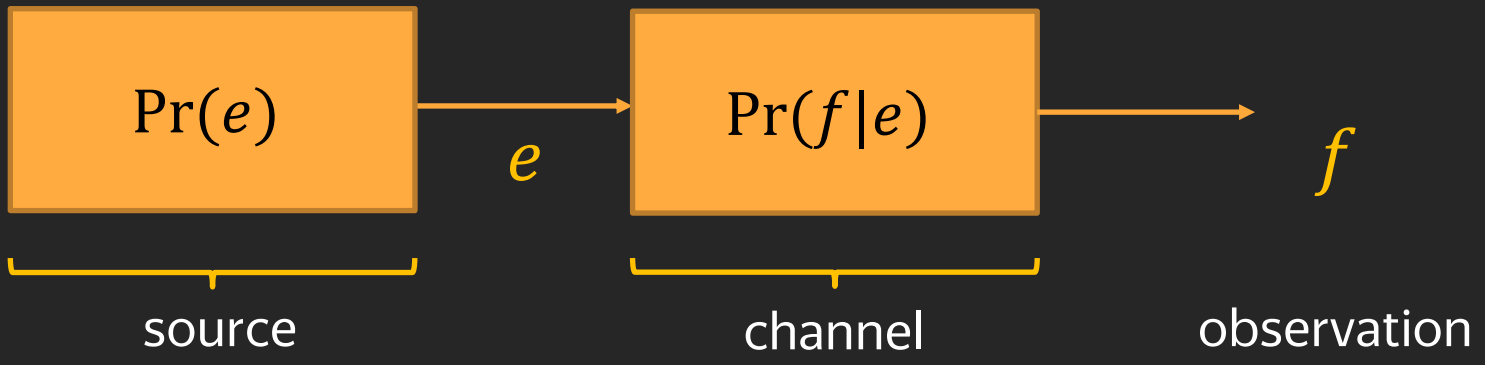


The Noisy Channel



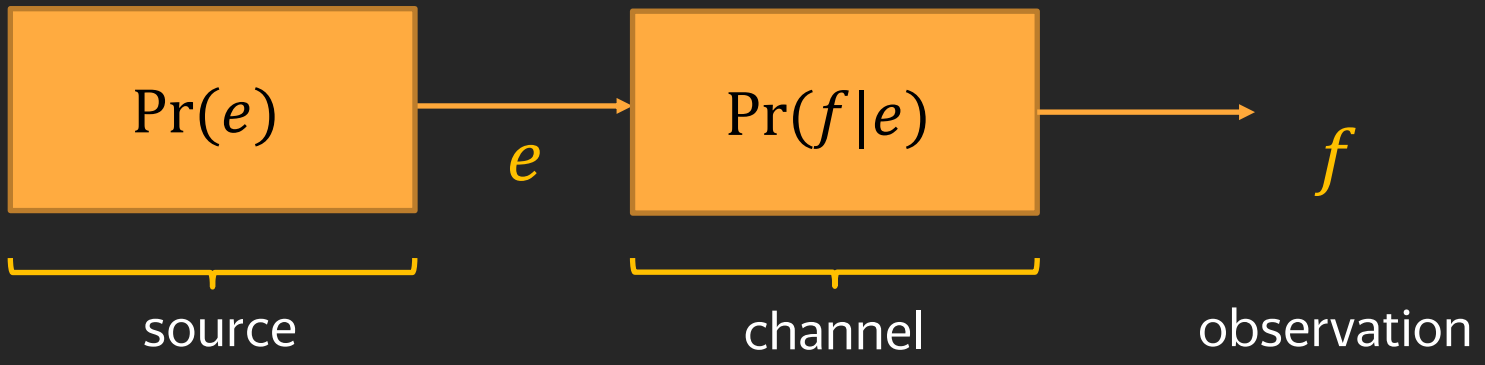
$$e^* = argmax_e \ Pr(e|f)$$

The Noisy Channel



$$e^* = argmax_e \; Pr(e) \; \ldots$$

The Noisy Channel



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

Advantages of Noisy Channel Decomposition

Advantages of Noisy Channel Decomposition

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

Advantages of Noisy Channel Decomposition

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

1. Abundance of monolingual data for estimating $Pr(e)$

Advantages of Noisy Channel Decomposition

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

1. Abundance of monolingual data for estimating $Pr(e)$
2. Approximations in $Pr(f|e)$ less damaging than approximations in $Pr(e|f)$

Advantages of Noisy Channel Decomposition

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

1. Abundance of monolingual data for estimating $Pr(e)$
2. Approximations in $Pr(f|e)$ less damaging than approximations in $Pr(e|f)$

E.g. if we model words in the sentence f as conditionally independent given e

Advantages of Noisy Channel Decomposition

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

1. Abundance of monolingual data for estimating $Pr(e)$
2. Approximations in $Pr(f|e)$ less damaging than approximations in $Pr(e|f)$

E.g. if we model words in the sentence f as conditionally independent given e

$$Pr(f|e) \approx \prod_{j=1}^{J} Pr(f_j|e)$$

Advantages of Noisy Channel Decomposition

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

1. Abundance of monolingual data for estimating $Pr(e)$
2. Approximations in $Pr(f|e)$ less damaging than approximations in $Pr(e|f)$

E.g. if we model words in the sentence f as *conditionally independent* given e

$$Pr(f|e) \approx \prod_{j=1}^{J} Pr(f_j|e)$$

the observations f are *not* independent at inference time when the sentence e is not observed

Advantages of Noisy Channel Decomposition

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

1. Abundance of monolingual data for estimating $Pr(e)$
2. Approximations in $Pr(f|e)$ less damaging than approximations in $Pr(e|f)$

E.g. if we model words in the sentence f as *conditionally independent* given e

$$Pr(f|e) \approx \prod_{j=1}^{J} Pr(f_j|e)$$

the observations f are *not* independent at inference time when the sentence e is not observed since each word f_j may provide information about e and therefore affect $Pr(f_j|e)$ for all other j.

Conditional independence

Conditional independence

- We have two coins (red and blue) with known probabilities of heads:

Conditional independence

- We have two coins (red and blue) with known probabilities of heads:

$$Pr(head|red) = 0.2, \qquad Pr(head|blue) = 0.8$$

Conditional independence

- We have two coins (red and blue) with known probabilities of heads:

$$Pr(head|red) = 0.2, \qquad Pr(head|blue) = 0.8$$

- If we take the red coin and flip it 5 times are the observations independent?

Conditional independence

- We have two coins (red and blue) with known probabilities of heads:

$$Pr(head|red) = 0.2, \qquad Pr(head|blue) = 0.8$$

- If we take the red coin and flip it 5 times are the observations independent?

Conditional independence

- We have two coins (red and blue) with known probabilities of heads:

$$Pr(head|red) = 0.2, \qquad Pr(head|blue) = 0.8$$

- If we take the red coin and flip it 5 times are the observations independent?

Conditional independence

- We have two coins (red and blue) with known probabilities of heads:

$$Pr(head|red) = 0.2, \qquad Pr(head|blue) = 0.8$$

- If we take the red coin and flip it 5 times are the observations independent?

Conditional independence

- We have two coins (red and blue) with known probabilities of heads:

$$Pr(head|red) = 0.2, \qquad Pr(head|blue) = 0.8$$

- If we take the red coin and flip it 5 times are the observations independent?

Conditional independence

- We have two coins (red and blue) with known probabilities of heads:

$$Pr(head|red) = 0.2, \qquad Pr(head|blue) = 0.8$$

- If we take the red coin and flip it 5 times are the observations independent?

Conditional independence

- We have two coins (red and blue) with known probabilities of heads:

$$Pr(head|red) = 0.2, \qquad Pr(head|blue) = 0.8$$

- If we take the red coin and flip it 5 times are the observations independent?

- If the colors are washed off the coins and we choose one at random are the observations independent?

Conditional independence

- We have two coins (red and blue) with known probabilities of heads:

$$Pr(head|red) = 0.2, \qquad Pr(head|blue) = 0.8$$

- If we take the red coin and flip it 5 times are the observations independent?

- If the colors are washed off the coins and we choose one at random are the observations independent?

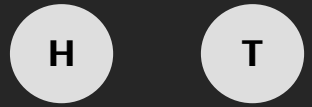
Conditional independence

- We have two coins (red and blue) with known probabilities of heads:

$$Pr(head|red) = 0.2, \qquad Pr(head|blue) = 0.8$$

- If we take the red coin and flip it 5 times are the observations independent?

- If the colors are washed off the coins and we choose one at random are the observations independent?



Conditional independence

- We have two coins (red and blue) with known probabilities of heads:

$$Pr(head|red) = 0.2, \qquad Pr(head|blue) = 0.8$$

- If we take the red coin and flip it 5 times are the observations independent?

- If the colors are washed off the coins and we choose one at random are the observations independent?

Conditional independence

- We have two coins (red and blue) with known probabilities of heads:

$$Pr(head|red) = 0.2, \qquad Pr(head|blue) = 0.8$$

- If we take the red coin and flip it 5 times are the observations independent?

- If the colors are washed off the coins and we choose one at random are the observations independent?

Conditional independence

- We have two coins (red and blue) with known probabilities of heads:

$$Pr(head|red) = 0.2, \qquad Pr(head|blue) = 0.8$$

- If we take the red coin and flip it 5 times are the observations independent?

- If the colors are washed off the coins and we choose one at random are the observations independent?

Conditional independence

- We have two coins (red and blue) with known probabilities of heads:

$$Pr(head|red) = 0.2, \qquad Pr(head|blue) = 0.8$$

- If we take the red coin and flip it 5 times are the observations independent?

- If the colors are washed off the coins and we choose one at random are the observations independent?

Conditional independence

- We have two coins (red and blue) with known probabilities of heads:

$$Pr(head|red) = 0.2, \qquad Pr(head|blue) = 0.8$$

- If we take the red coin and flip it 5 times are the observations independent?

- If the colors are washed off the coins and we choose one at random are the observations independent?

Advantages of Noisy Channel Decomposition

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

Advantages of Noisy Channel Decomposition

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

- Independence assumptions help reduce parameter space

Advantages of Noisy Channel Decomposition

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

- Independence assumptions help reduce parameter space
- Independence assumptions help simplify inference (computation)

Advantages of Noisy Channel Decomposition

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

- Independence assumptions help reduce parameter space
- Independence assumptions help simplify inference (computation)
- Directly maximizing $Pr(e|f)$ with same factorization would be awful, e.g.

Advantages of Noisy Channel Decomposition

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

- Independence assumptions help reduce parameter space
- Independence assumptions help simplify inference (computation)
- Directly maximizing $Pr(e|f)$ with same factorization would be awful, e.g.

$$\begin{aligned} e^* &= argmax_e \; Pr(e|f) \\ &\approx argmax_{e_i} \; \prod_{i=1}^{I} Pr(e_i|f) \end{aligned}$$

Advantages of Noisy Channel Decomposition

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

- Independence assumptions help reduce parameter space
- Independence assumptions help simplify inference (computation)
- Directly maximizing $Pr(e|f)$ with same factorization would be awful, e.g.

$$e^* = argmax_e \; Pr(e|f)$$
$$\approx argmax_{e_i} \prod_{i=1}^{I} Pr(e_i|f)$$

Advantages of Noisy Channel Decomposition

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

- Independence assumptions help reduce parameter space
- Independence assumptions help simplify inference (computation)
- Directly maximizing $Pr(e|f)$ with same factorization would be awful, e.g.

$$e^* = argmax_e \; Pr(e|f)$$
$$\approx argmax_{e_i} \prod_{i=1}^{I} Pr(e_i|f)$$

- With source-channel decomposition, dependencies are still modelled

Advantages of Noisy Channel Decomposition

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

- Independence assumptions help reduce parameter space
- Independence assumptions help simplify inference (computation)
- Directly maximizing $Pr(e|f)$ with same factorization would be awful, e.g.

$$e^* = argmax_e \; Pr(e|f)$$
$$\approx argmax_{e_i} \prod_{i=1}^{I} Pr(e_i|f)$$

- With source-channel decomposition, dependencies are still modelled

$$e^* = argmax_e \; Pr(e)\, Pr(f|e)$$
$$\approx argmax_e \; Pr(e) \prod_{j=1}^{J} Pr(f_j|e)$$

Word Alignment

Factorize the probability of observed sentence f given e into probability of generating each word f_j given only the words it's aligned to e_{a_j}, i.e.

Word Alignment

Factorize the probability of observed sentence f given e into probability of generating each word f_j given only the words it's aligned to e_{a_j}, i.e.

$$Pr(f|e,a) \approx \prod_{j=1}^{J} Pr\left(f_j \middle| e_{a_j}\right).$$

Word Alignment

Factorize the probability of observed sentence f given e into probability of generating each word f_j given only the words it's aligned to e_{a_j}, i.e.

$$Pr(f|e,a) \approx \prod_{j=1}^{J} Pr\left(f_j \middle| e_{a_j}\right).$$

The animal didn't cross the road because it was too tired

Животное не прошло дорогу потому что оно был слишком устало

Word Alignment

Learn parameters for each pair of distinct English and French words

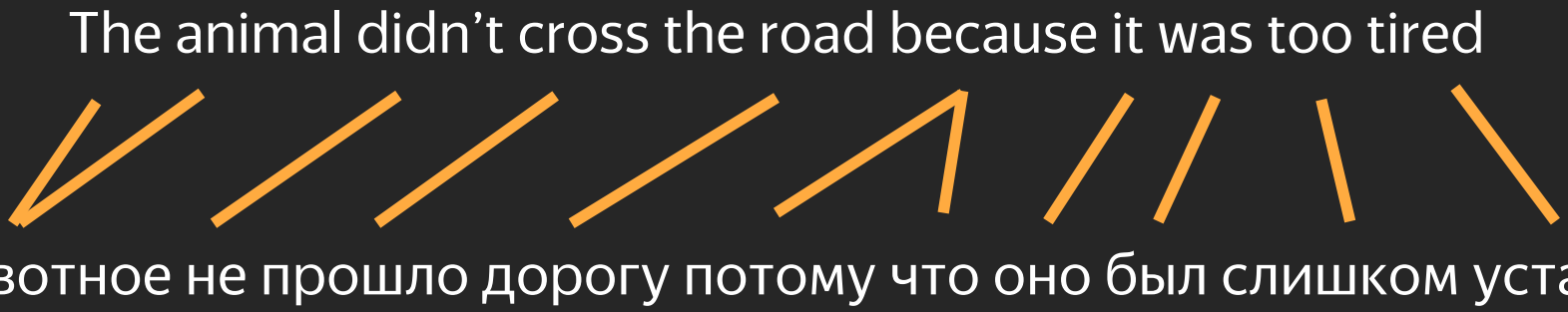
The animal didn’t cross the road because it was too tired

Животное не прошло дорогу потому что оно был слишком устало

Word Alignment

Learn parameters for each pair of distinct English and French words, i.e. let

$$Pr(f_j | e_i) = t(f|e).$$



Word Alignment

Learn parameters for each pair of distinct English and French words, i.e. let

$$Pr(f_j | e_i) = t(f|e).$$

E.g. $t(\text{животное}|\text{animal})$, $t(\text{прошло}|\text{cross})$, etc.

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

Животное не прошло дорогу потому что оно был слишком устало

Expectation Maximization Algorithm

Expectation Maximization Algorithm

Goal: Learn word level translation probabilities $t(f|e)$

Expectation Maximization Algorithm

Goal: Learn word level translation probabilities $t(f|e)$
Problem: Word level alignments are not observed

Expectation Maximization Algorithm

Goal: Learn word level translation probabilities $t(f|e)$
Problem: Word level alignments are not observed

1. Make an initial guess as to the word level parameters $t(f|e)$

Expectation Maximization Algorithm

Goal: Learn word level translation probabilities $t(f|e)$
Problem: Word level alignments are not observed

1. Make an initial guess as to the word level parameters $t(f|e)$
2. Infer the posterior distribution over alignments $Pr(a|e,f)$ given $t(f|e)$

Expectation Maximization Algorithm

Goal: Learn word level translation probabilities $t(f|e)$
Problem: Word level alignments are not observed

1. Make an initial guess as to the word level parameters $t(f|e)$
2. Infer the posterior distribution over alignments $Pr(a|e,f)$ given $t(f|e)$
3. Treat this posterior distribution as ‘fractional’ counts of alignments

Expectation Maximization Algorithm

Goal: Learn word level translation probabilities $t(f|e)$
Problem: Word level alignments are not observed

1. Make an initial guess as to the word level parameters $t(f|e)$
2. Infer the posterior distribution over alignments $Pr(a|e,f)$ given $t(f|e)$
3. Treat this posterior distribution as ‘fractional’ counts of alignments
4. Re-estimate word level parameters $t(f|e)$

Expectation Maximization Algorithm

Goal: Learn word level translation probabilities $t(f|e)$
Problem: Word level alignments are not observed

1. Make an initial guess as to the word level parameters $t(f|e)$
2. Infer the posterior distribution over alignments $Pr(a|e,f)$ given $t(f|e)$
3. Treat this posterior distribution as ‘fractional’ counts of alignments
4. Re-estimate word level parameters $t(f|e)$
5. Return to 2. until parameters $t(f|e)$ converge

Expectation Maximization Algorithm

Goal: Learn word level translation probabilities $t(f|e)$
Problem: Word level alignments are not observed

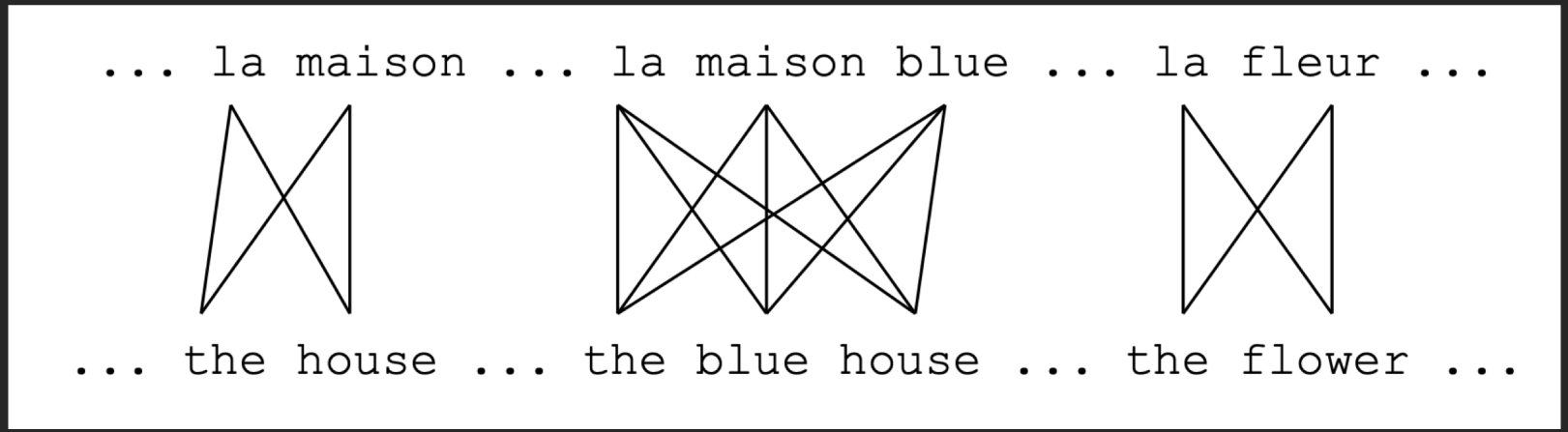
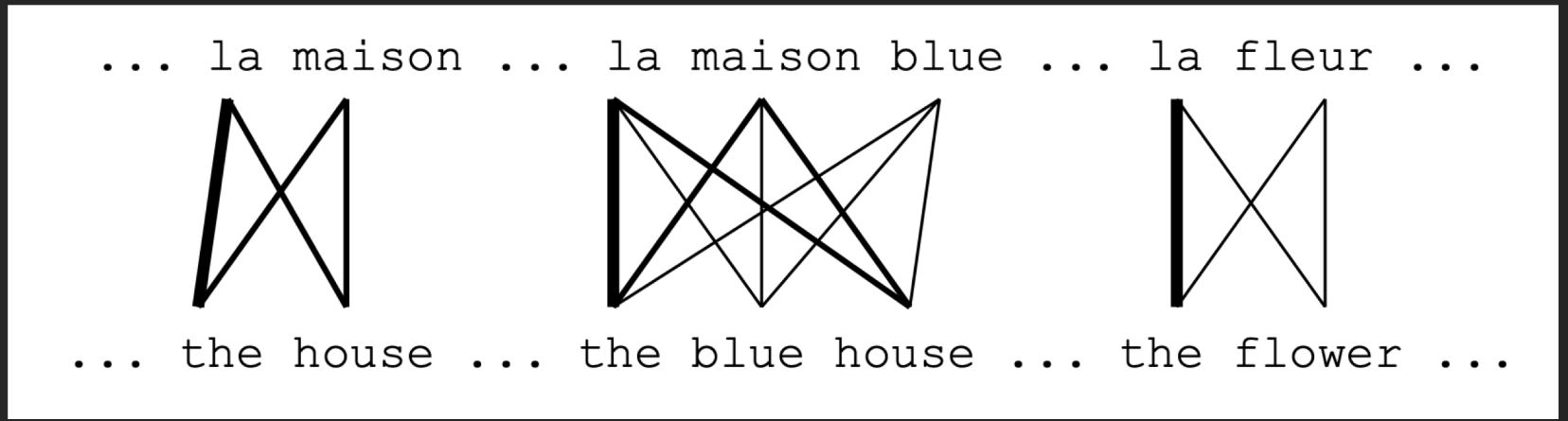


Image by Philipp Koehn

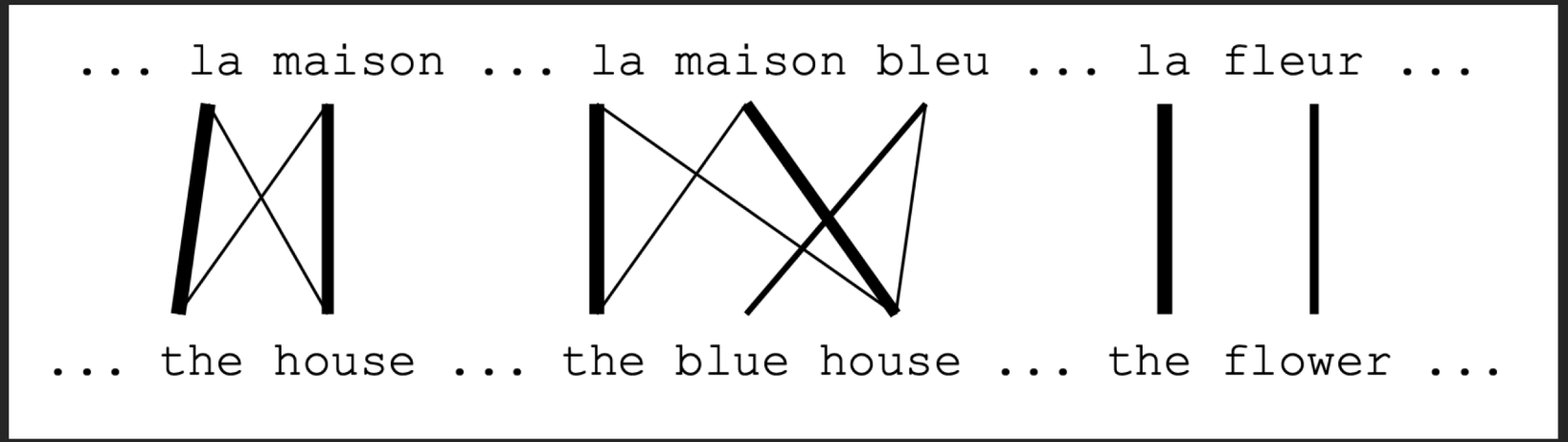
Expectation Maximization Algorithm

Goal: Learn word level translation probabilities $t(f|e)$
Problem: Word level alignments are not observed



Expectation Maximization Algorithm

Goal: Learn word level translation probabilities $t(f|e)$
Problem: Word level alignments are not observed



Expectation Maximization Algorithm

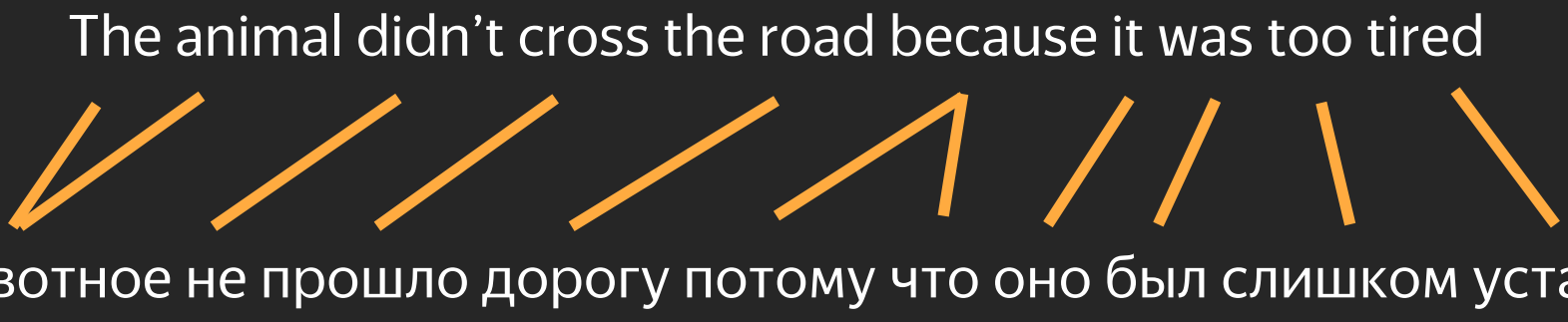
Goal: Learn word level translation probabilities $t(f|e)$
Problem: Word level alignments are not observed

Phrase-based Extensions (Koehn 2003)

Build phrasal models on top of word alignments

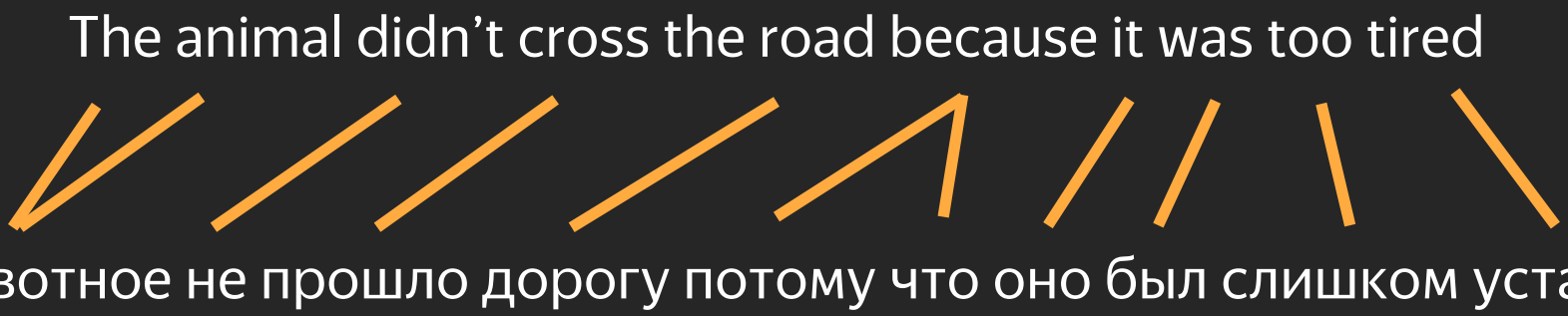
Phrase-based Extensions (Koehn 2003)

Build phrasal models on top of word alignments



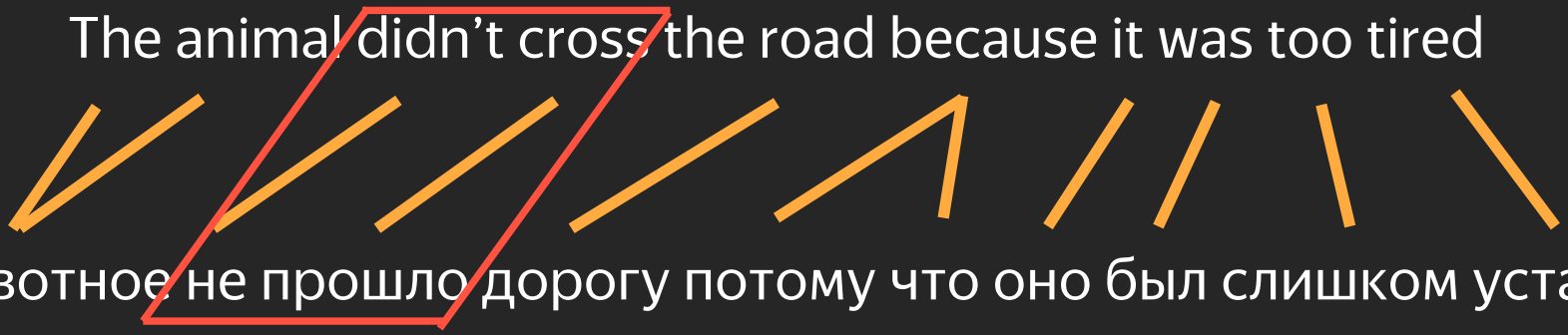
Phrase-based Extensions (Koehn 2003)

Build phrasal models on top of word alignments
Extract arbitrary phrases that are *more-or-less* mutually aligned



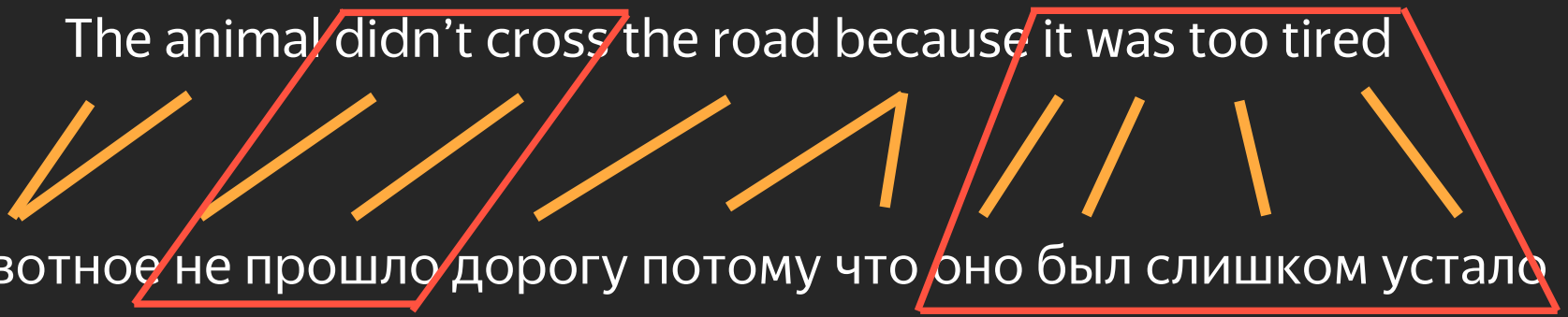
Phrase-based Extensions (Koehn 2003)

Build phrasal models on top of word alignments
Extract arbitrary phrases that are *more-or-less* mutually aligned



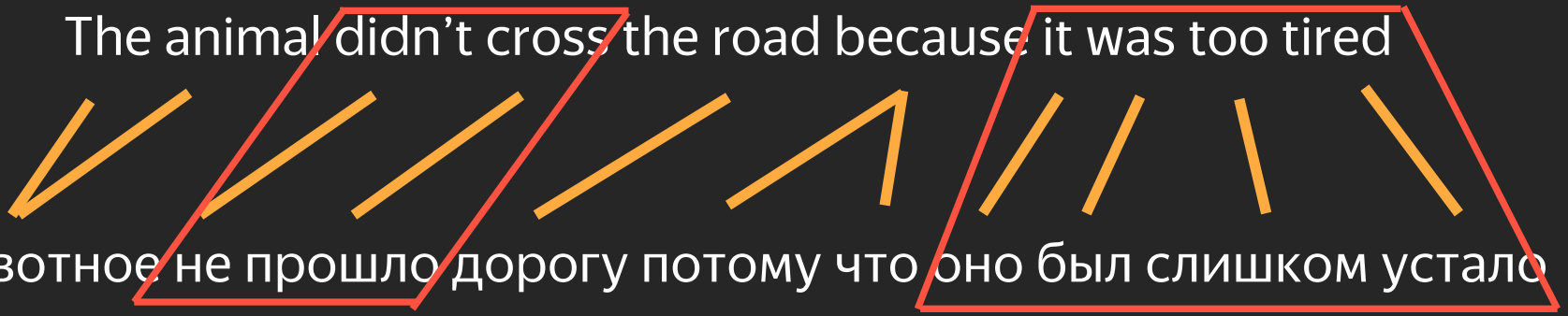
Phrase-based Extensions (Koehn 2003)

Build phrasal models on top of word alignments
Extract arbitrary phrases that are *more-or-less* mutually aligned



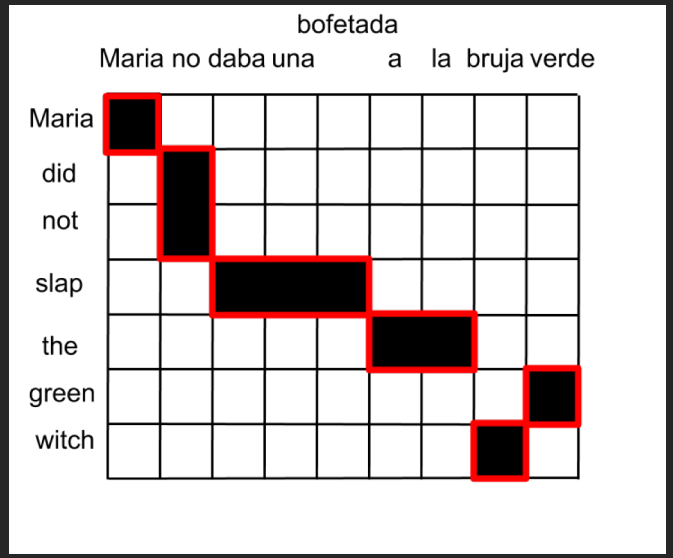
Phrase-based Extensions (Koehn 2003)

Build phrasal models on top of word alignments
Extract arbitrary phrases that are *more-or-less* mutually aligned
Estimate *phrase-level* translation probabilities

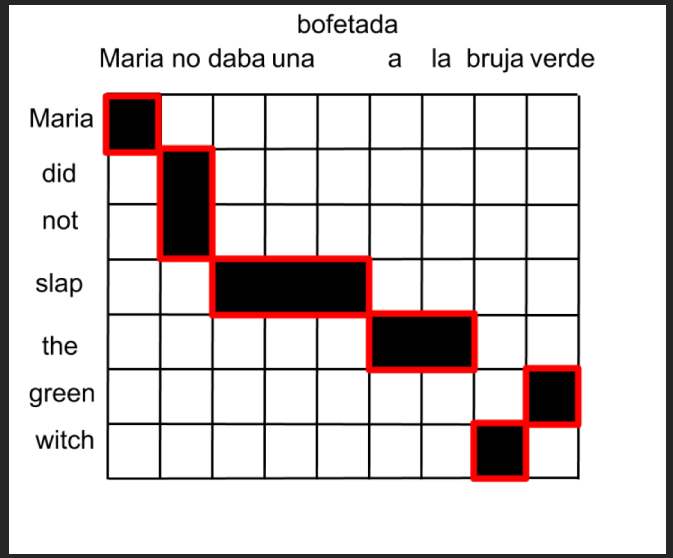


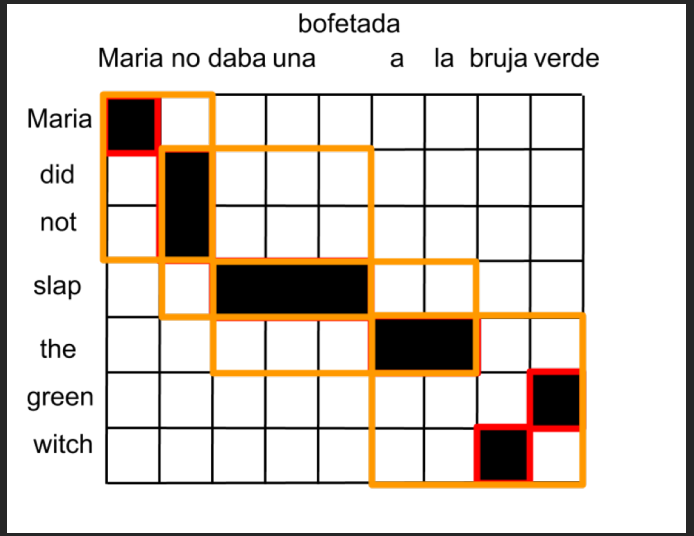
From Words to Phrases

From Words to Phrases



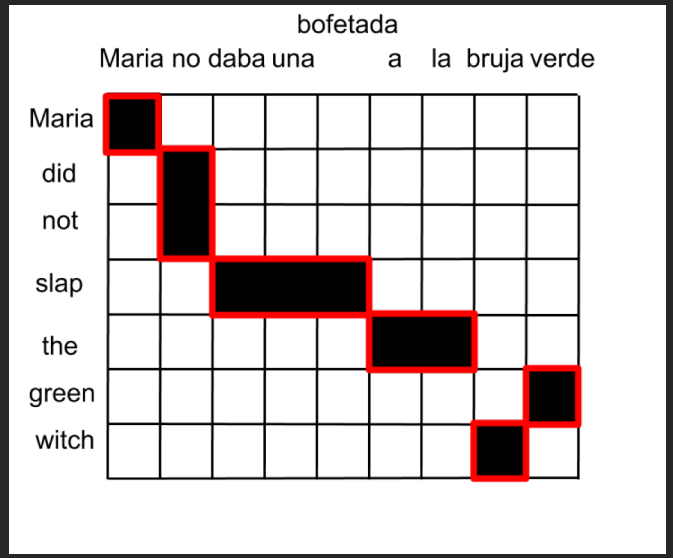
From Words to Phrases

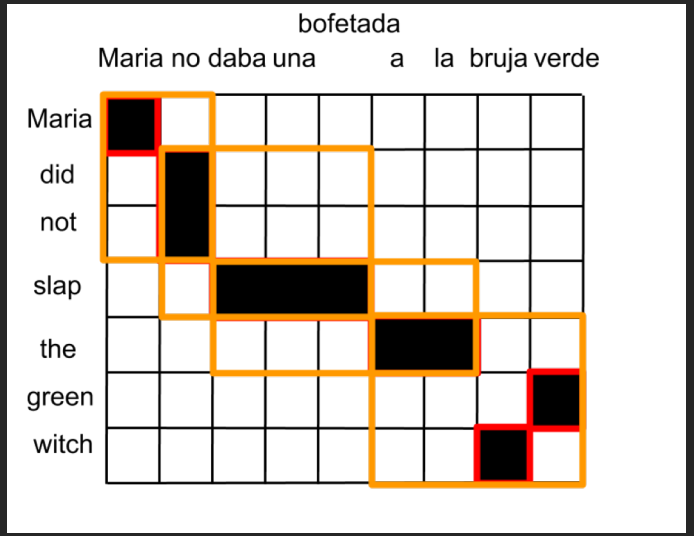




From Words to Phrases

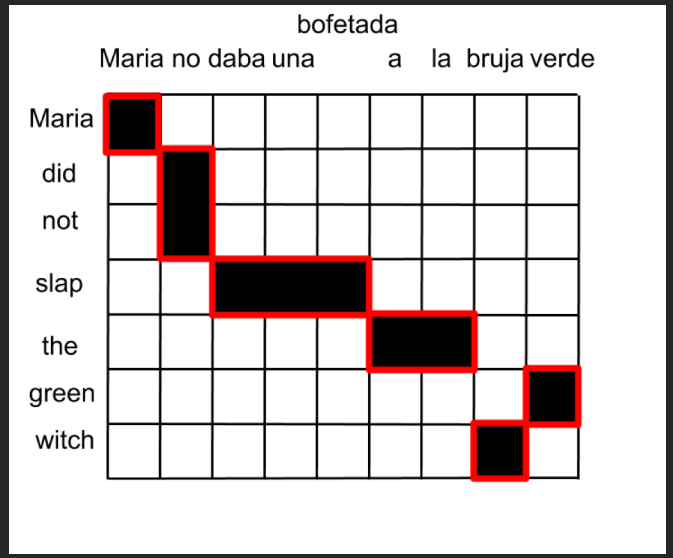
- Estimate word alignments using EM

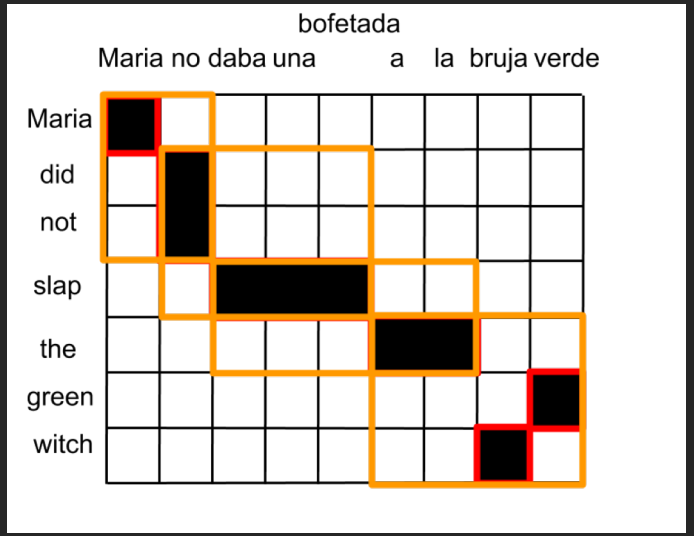




From Words to Phrases

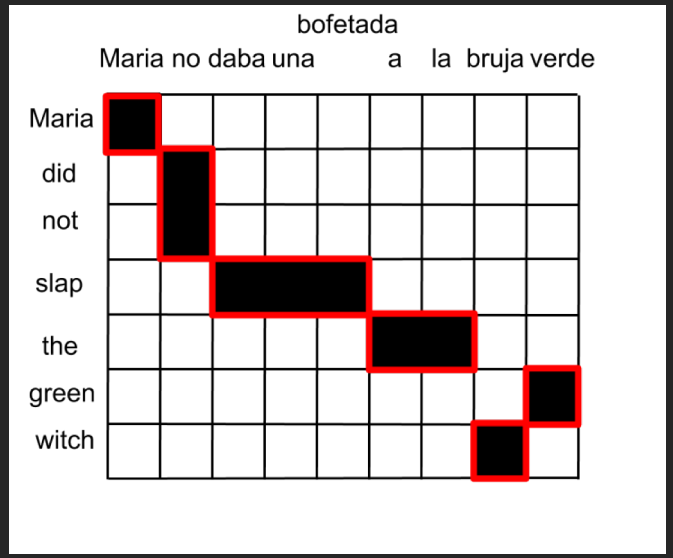
- Estimate word alignments using EM
- Use word alignments as constraints to align phrases

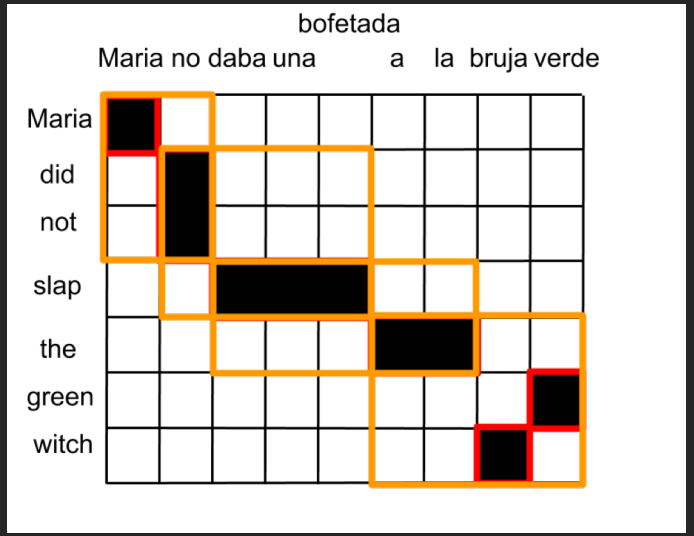




From Words to Phrases

- Estimate word alignments using EM
- Use word alignments as constraints to align phrases
- Build phrasal model of $\Pr(f|e)$





Phrasal Translation Model

Phrasal Translation Model

- Score phrase pairs based on counts of aligned phrase pairs

Phrasal Translation Model

- Score phrase pairs based on counts of aligned phrase pairs
- Add word level scores to smooth these

Phrasal Translation Model

- Score phrase pairs based on counts of aligned phrase pairs
- Add word level scores to smooth these
- Add arbitrary features to phrase, e.g. $Pr(e|f)$ in addition to $Pr(f|e)$

Phrase-based Translation Model

He	→	Он
stood	→	стоял, стояла, поставил, …
bank	→	берега, берегу, банк, банка …
He stood	→	Он стоял
by the bank	→	на берегу, рядом с банком …

Linear Translation Model

Linear Translation Model

- Arbitrary features $\phi_k(e, f)$: phrase-table, language model, length penalty, reordering costs, word-sense disambiguation, etc.

Linear Translation Model

- Arbitrary features $\phi_k(e, f)$: phrase-table, language model, length penalty, reordering costs, word-sense disambiguation, etc.

$$e^* = argmax_e \Sigma_k \lambda_k \phi_k(e, f)$$

Linear Translation Model

- Arbitrary features $\phi_k(e, f)$: phrase-table, language model, length penalty, reordering costs, word-sense disambiguation, etc.
- Build discriminative model using generative models as features

$$e^* = argmax_e \Sigma_k \lambda_k \phi_k(e, f)$$

Linear Translation Model

- Arbitrary features $\phi_k(e, f)$: phrase-table, language model, length penalty, reordering costs, word-sense disambiguation, etc.
- Build discriminative model using generative models as features
- Optimize evaluation metric (BLEU) directly with beam search on dev

$$e^* = argmax_e \Sigma_k \lambda_k \phi_k(e, f)$$

Phrase-based Decoder

Phrase-based Decoder

Goal: Find the highest scoring translation **that translated all the input**

Phrase-based Decoder

Goal: Find the highest scoring translation **that translated all the input**

Solution: Stack based decoding (beam search)

Phrase-based Decoder

Goal: Find the highest scoring translation **that translated all the input**

Solution: Stack based decoding (beam search)

- Start with an empty hypothesis

Phrase-based Decoder

Goal: Find the highest scoring translation **that translated all the input**

Solution: Stack based decoding (beam search)

- Start with an empty hypothesis
- Extend hypotheses by translating some (still) untranslated source words

Phrase-based Decoder

Goal: Find the highest scoring translation **that translated all the input**

Solution: Stack based decoding (beam search)

- Start with an empty hypothesis
- Extend hypotheses by translating some (still) untranslated source words
- Backtrack from highest scoring hypothesis that translates all words

Phrase-based Translation

Phrase-based Translation

English

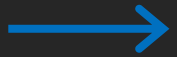
- He stood on the bank

Phrase-based Translation

English

- He stood on the bank

Phrase Table

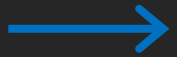


Phrase-based Translation

English *Pieces of Russian*

- He stood on the bank

Phrase Table

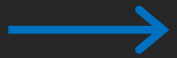


Phrase-based Translation

English *Pieces of Russian*

- He stood on the bank

Phrase Table



{Он | стояла | на | берегу ... }
Он | стоял | у | банка, ... }
Он | поставил | у берега, ... }

Phrase-based Translation

English *Pieces of Russian*

- He stood on the bank

Phrase Table → Language Model →

{Он | стояла | на | берегу ... }
Он | стоял | у | банка, ... }
Он | поставил | у берега, ... }

Phrase-based Translation

English *Pieces of Russian* *Russian*

- He stood on the bank Он стоял на берегу

Phrase Table → Language Model →

{Он | стояла | на | берегу ... }
Он | стоял | у | банка, ... }
Он | поставил | у берега, ... }

Phrase-based Extensions (Koehn 2003)

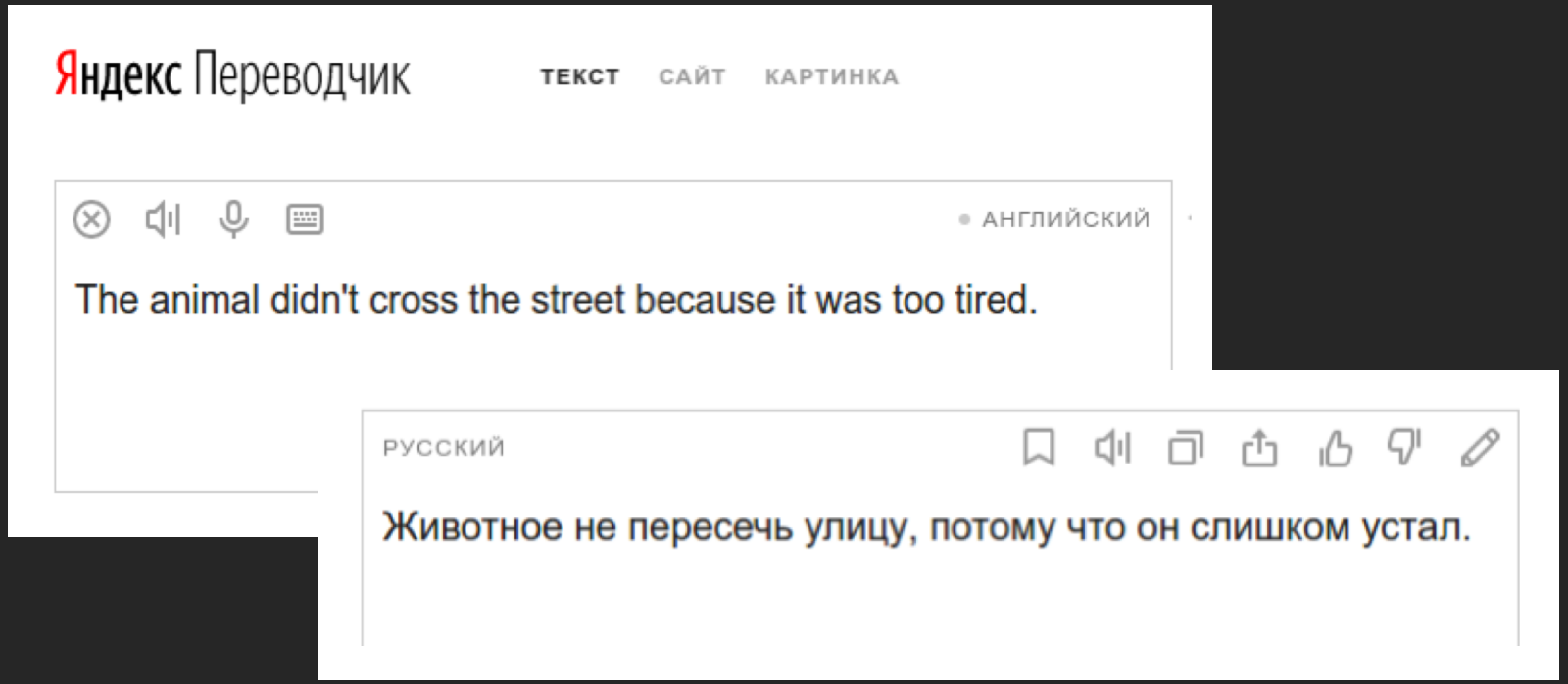
Phrase-based MT together with internet-scale parallel data mining enabled the first popular MT engines from mid-2000s

Яндекс Переводчик

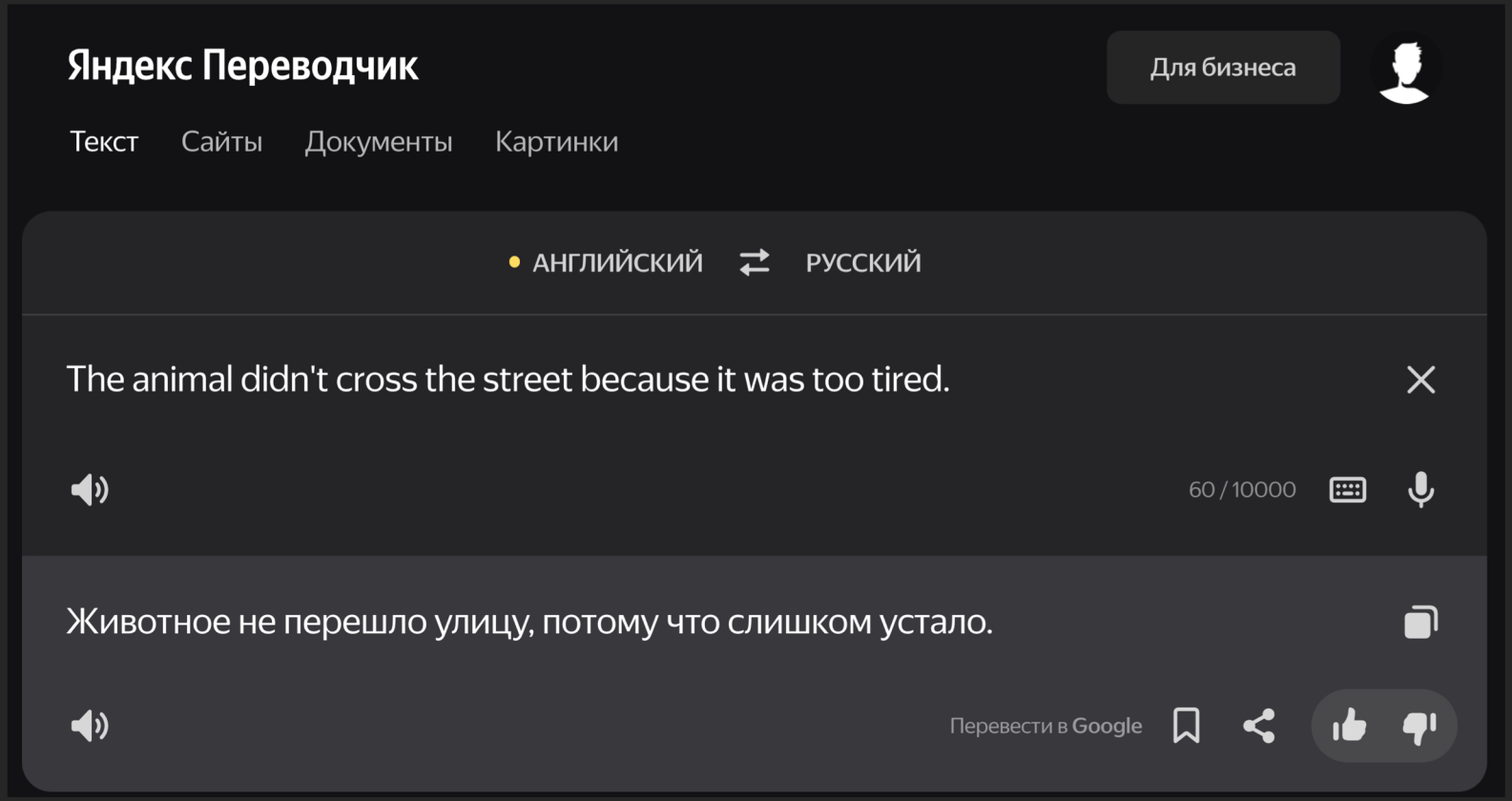
PROMT®

Phrase-based MT Quality (around 2016)

Great for getting the ‘gist’, not so great for sounding fluent or natural

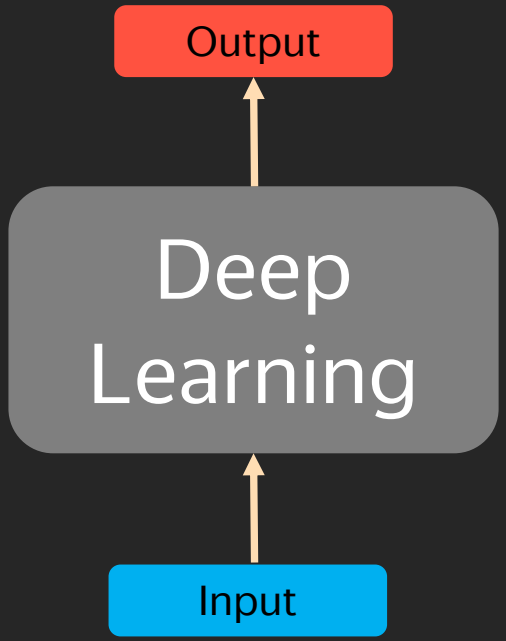


Something happened with MT Quality around 2017...

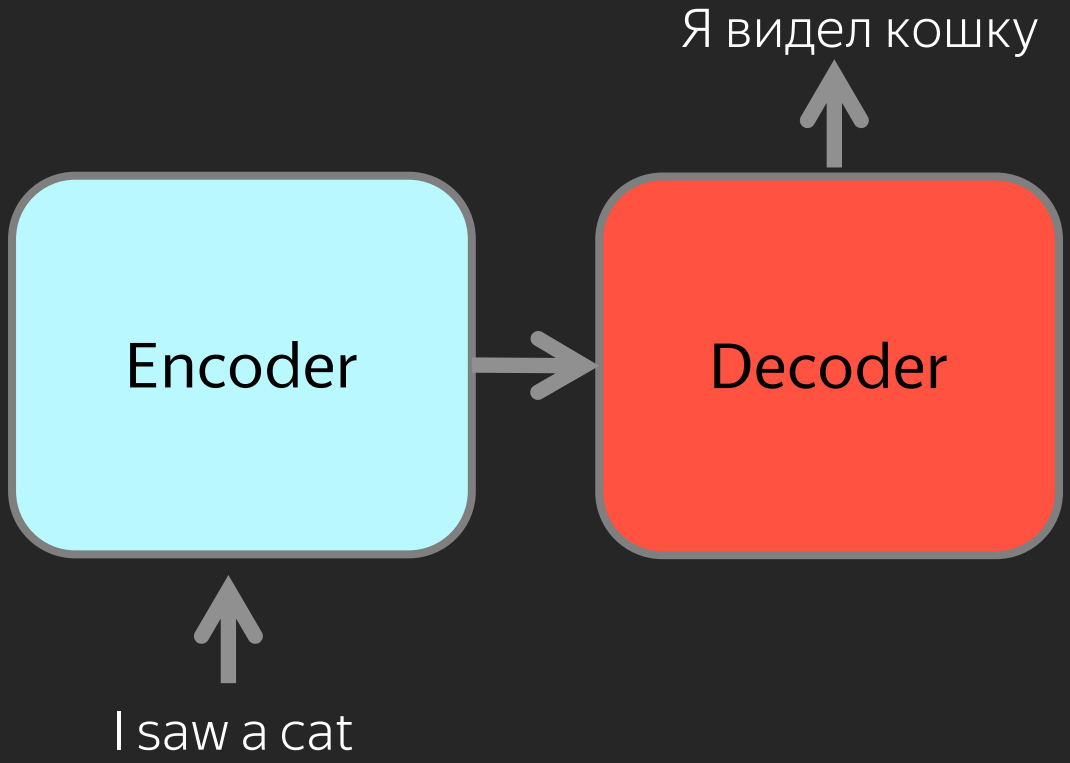


How Neural Machine Translation changed things

Neural Machine Translation



Encoder-Decoder



End-to-End Optimization

End-to-End Optimization

- Direct optimization of conditional probability of translation in corpus

$$Pr(e|f)$$

End-to-End Optimization

- Direct optimization of conditional probability of translation in corpus

$$Pr(e|f)$$

- Optimize conditional probability of next word (locally normalized)

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

End-to-End Optimization

- Direct optimization of conditional probability of translation in corpus

$$Pr(e|f)$$

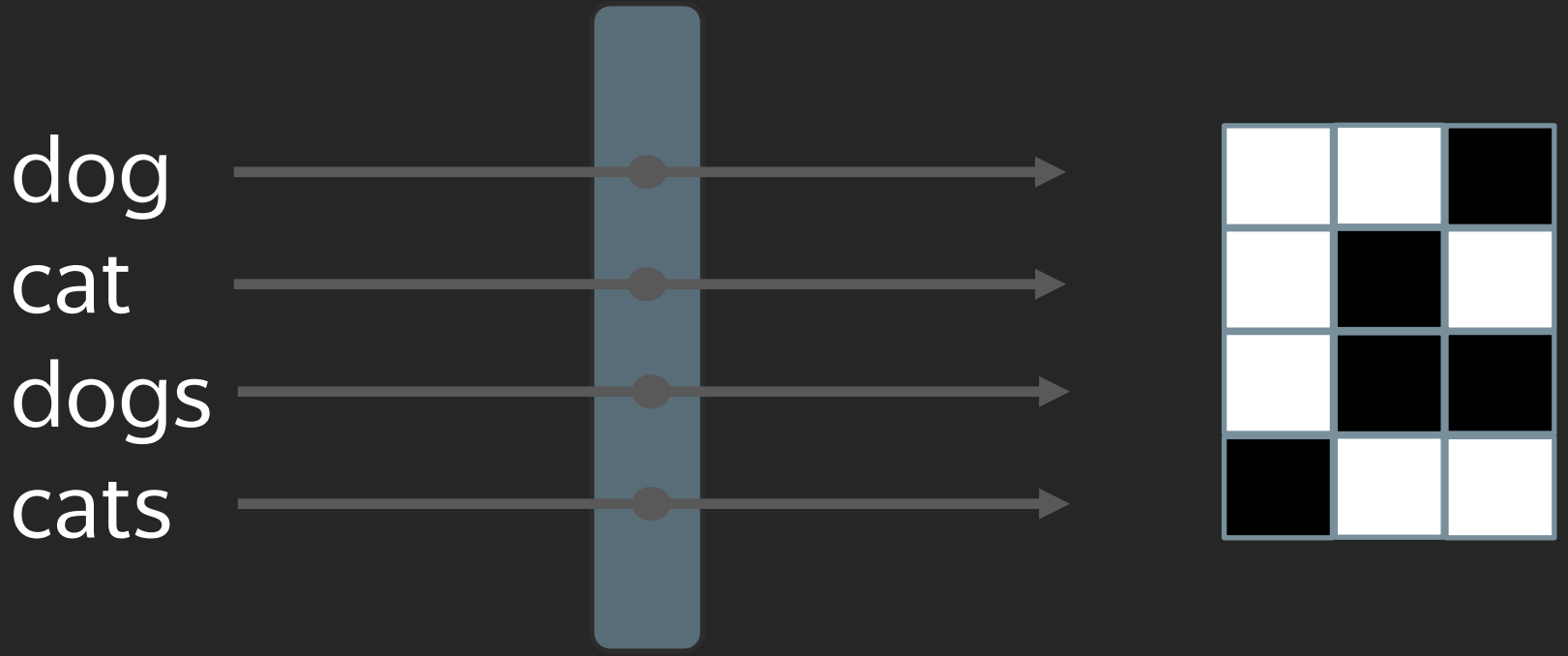
- Optimize conditional probability of next word (locally normalized)

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

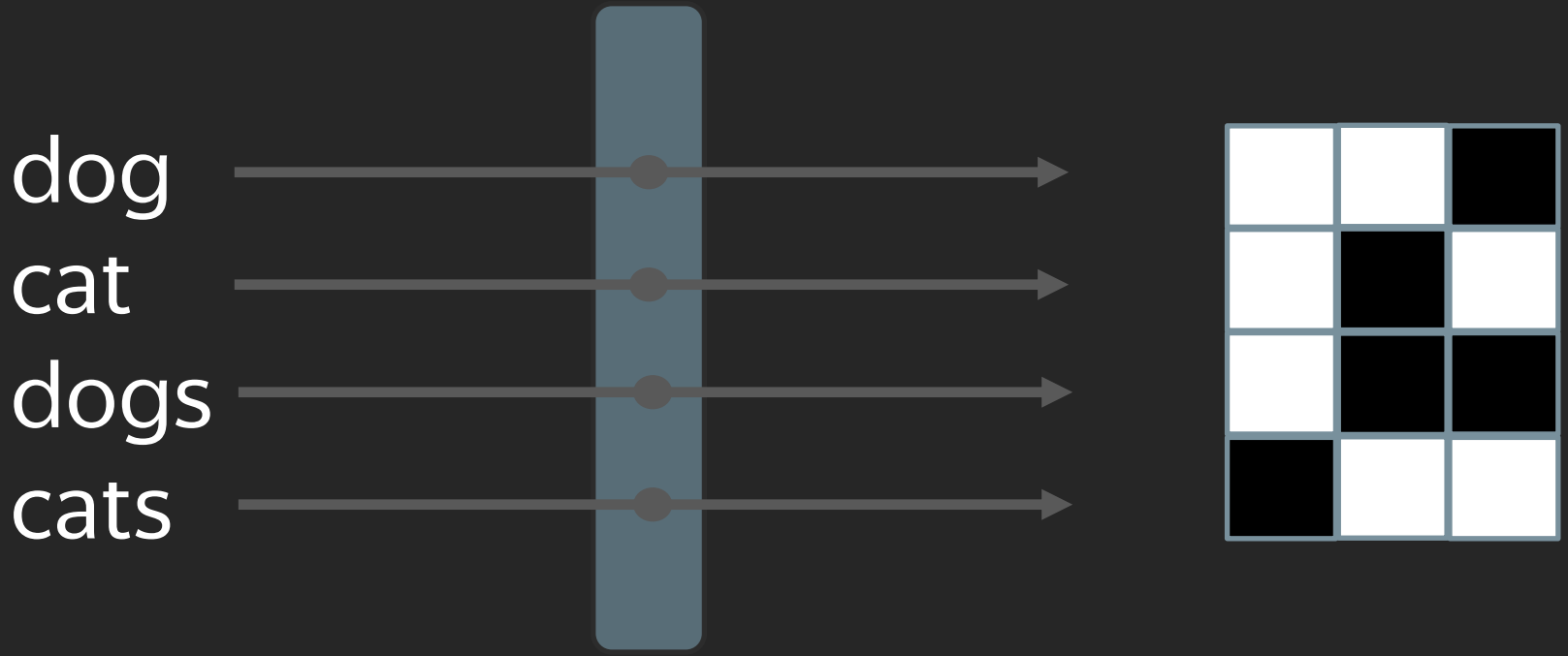
- Teacher forcing (i.e. provide reference prefix to model during training)

Representation Learning

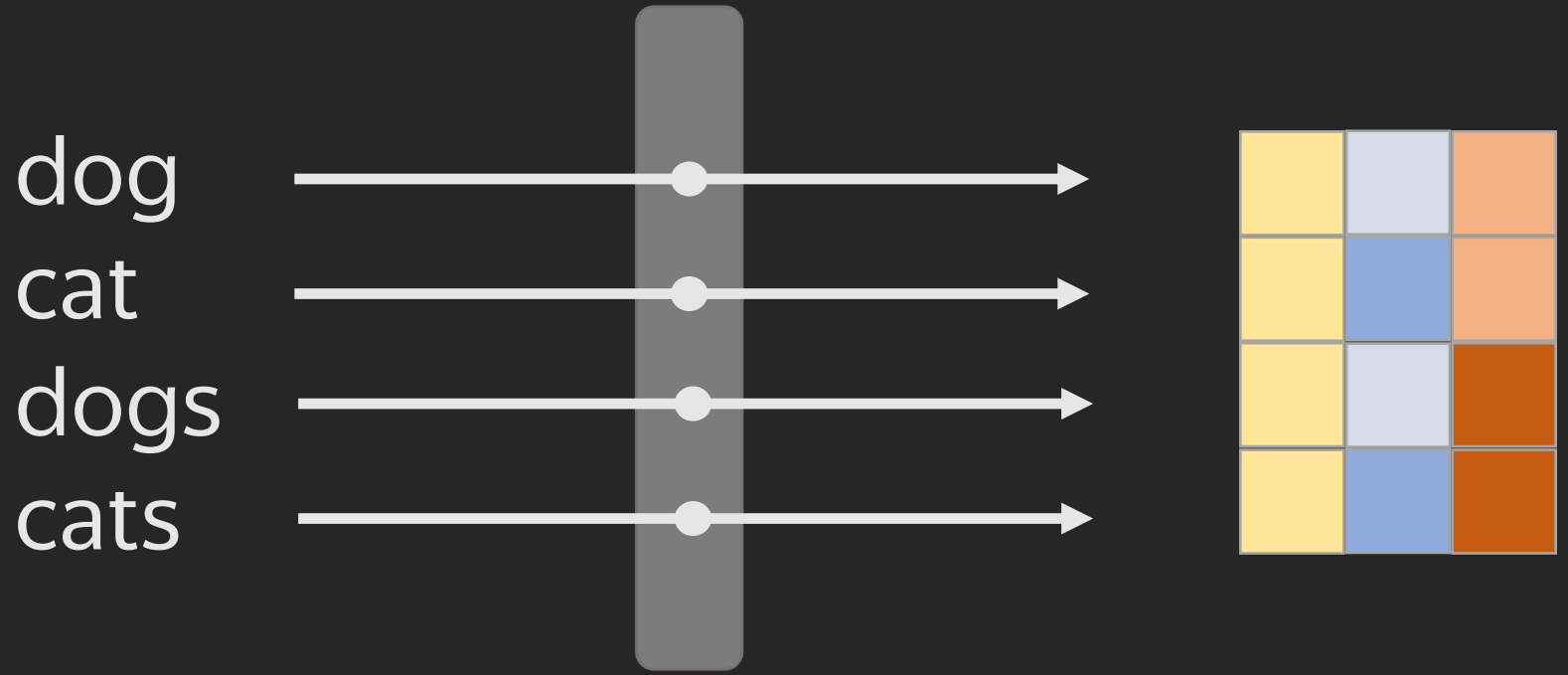
Non-learned Representations



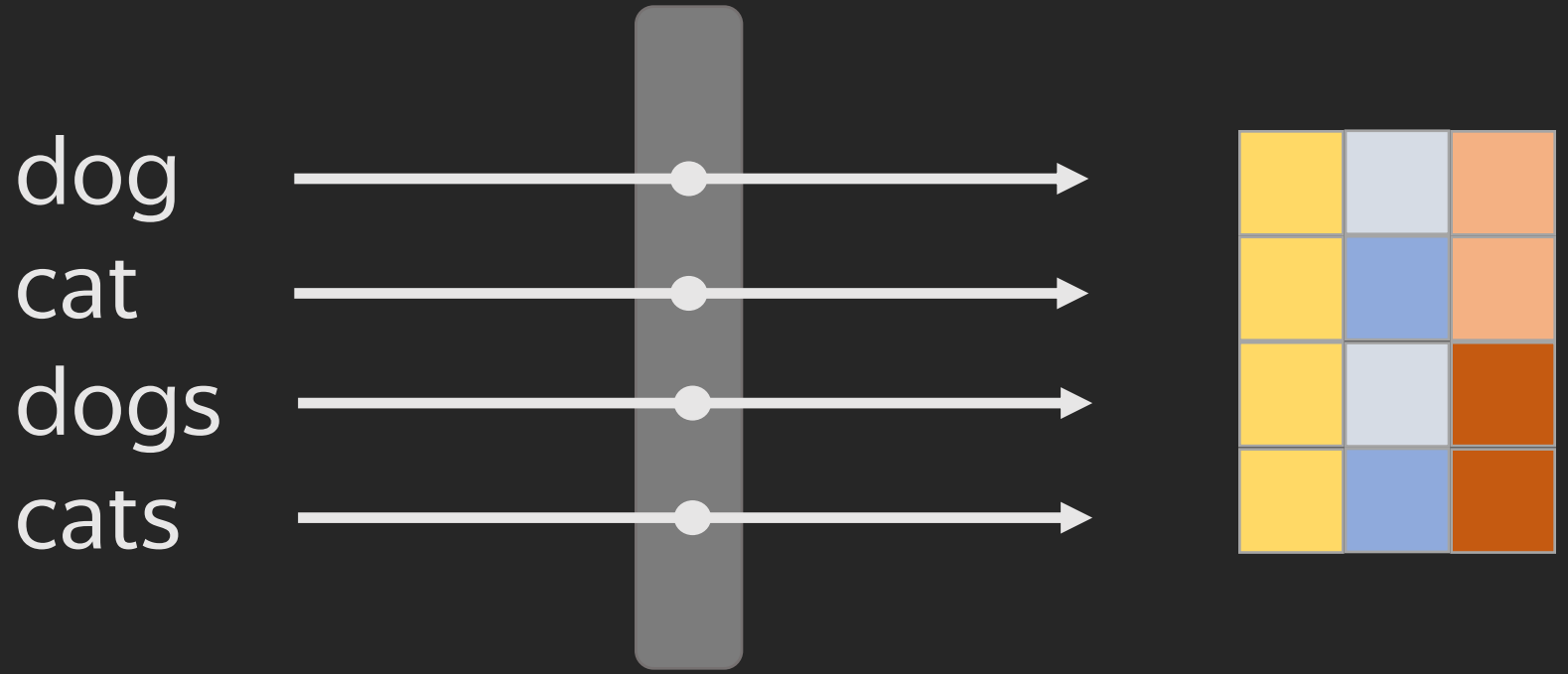
Non-learned Representations (Hamming Space)



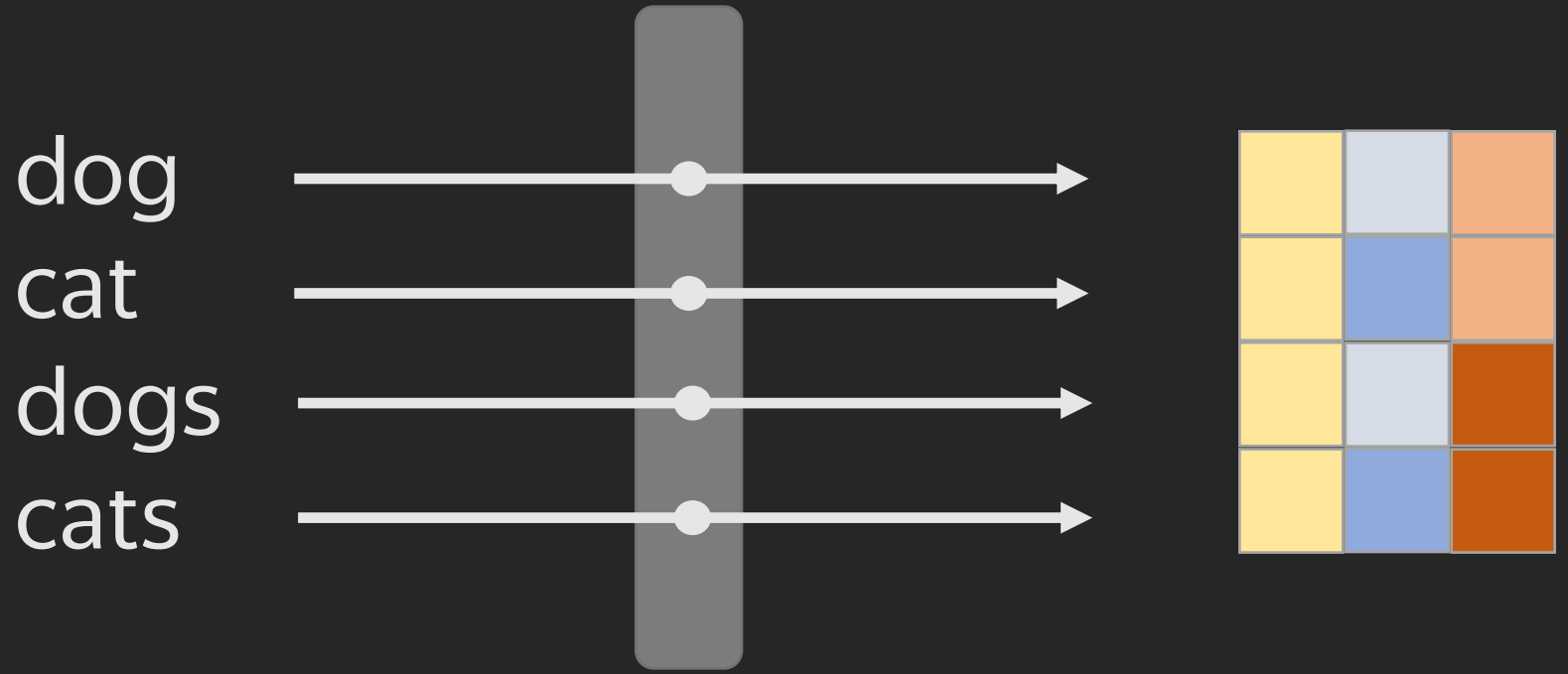
Representation Learning



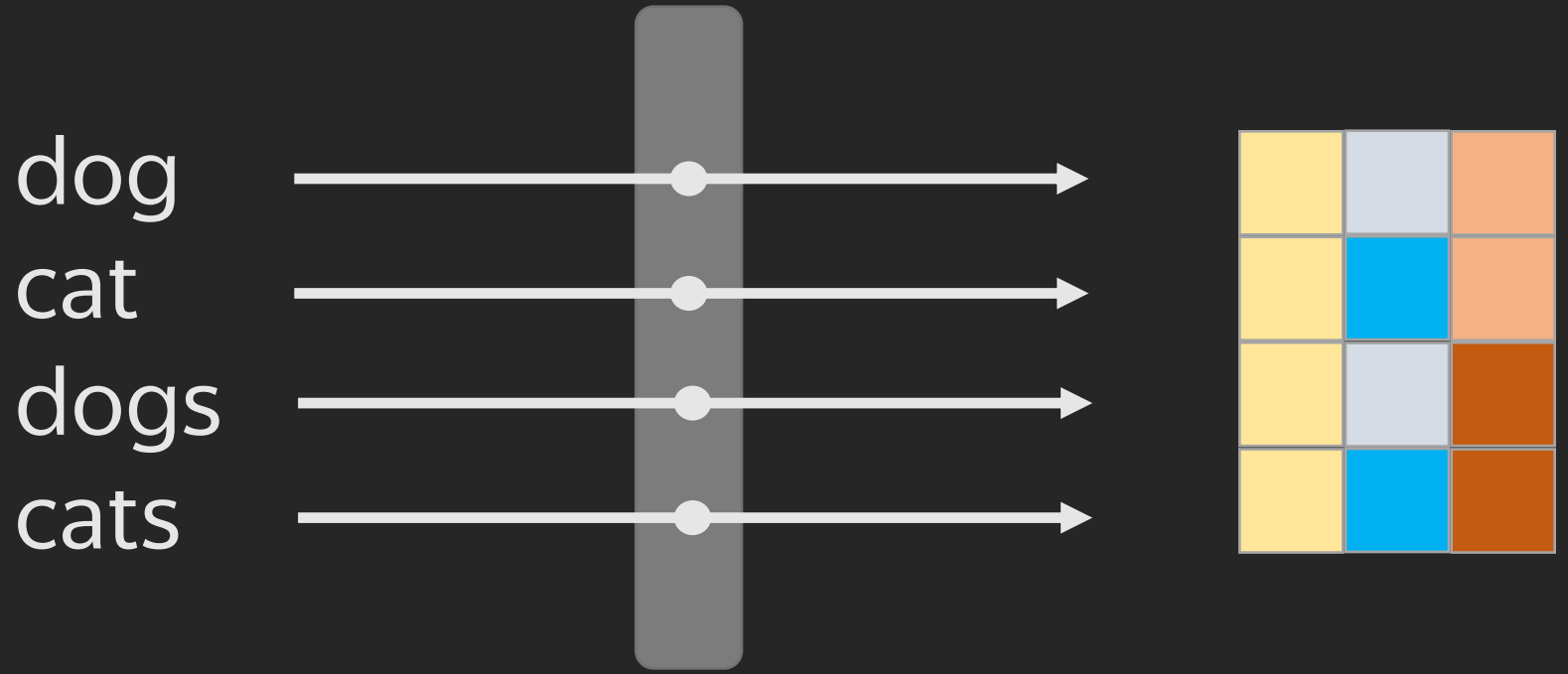
Representation Learning (Embedding space)



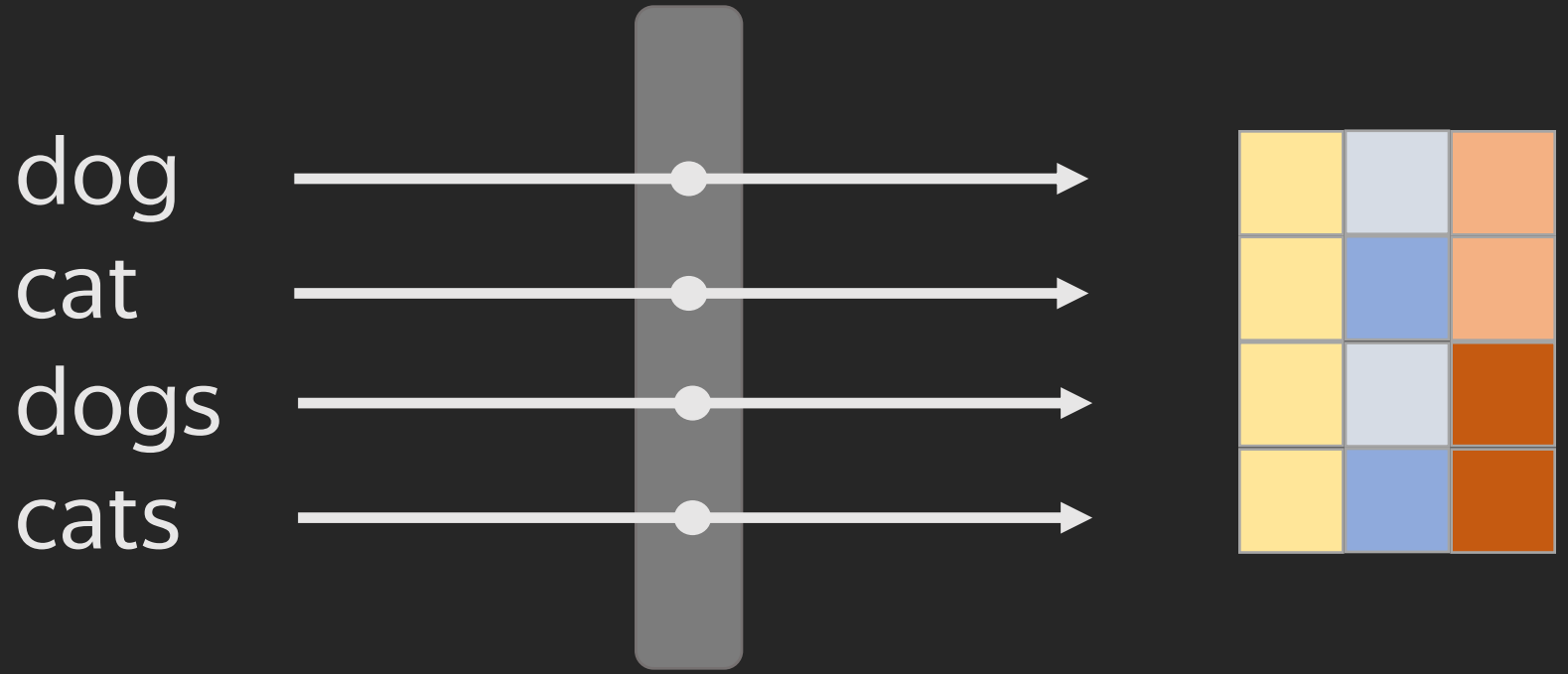
Representation Learning (Embedding space)



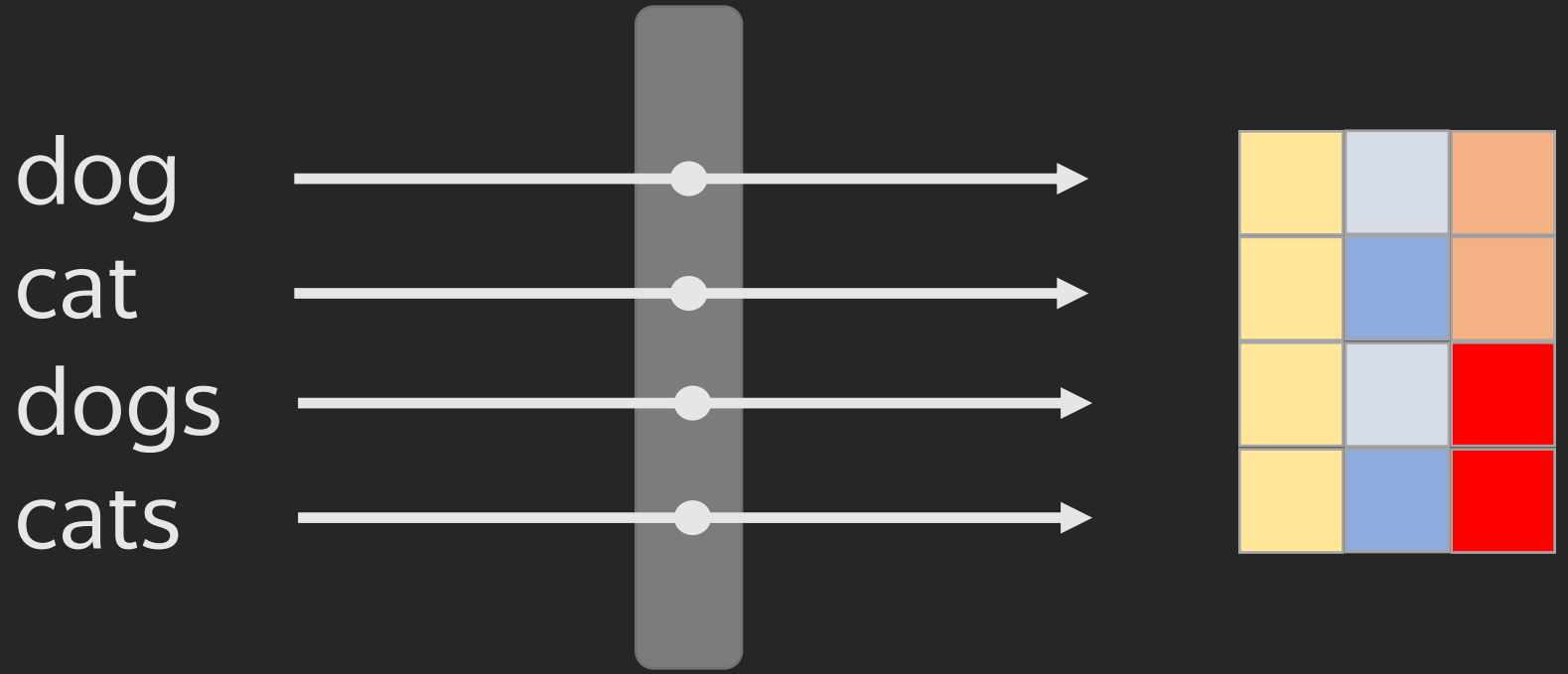
Representation Learning (Embedding space)



Representation Learning (Embedding space)



Representation Learning (Embedding space)



Independence Assumptions

- Chain-rule factorization (no explicit independence assumptions)

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

Conditional language model

Conditional language model

- Much more fluent output

Conditional language model

- Much more fluent output

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

Conditional language model

- Much more fluent output

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

PBMT: Животное не пересечь улицу потому, что он слишком устал

Conditional language model

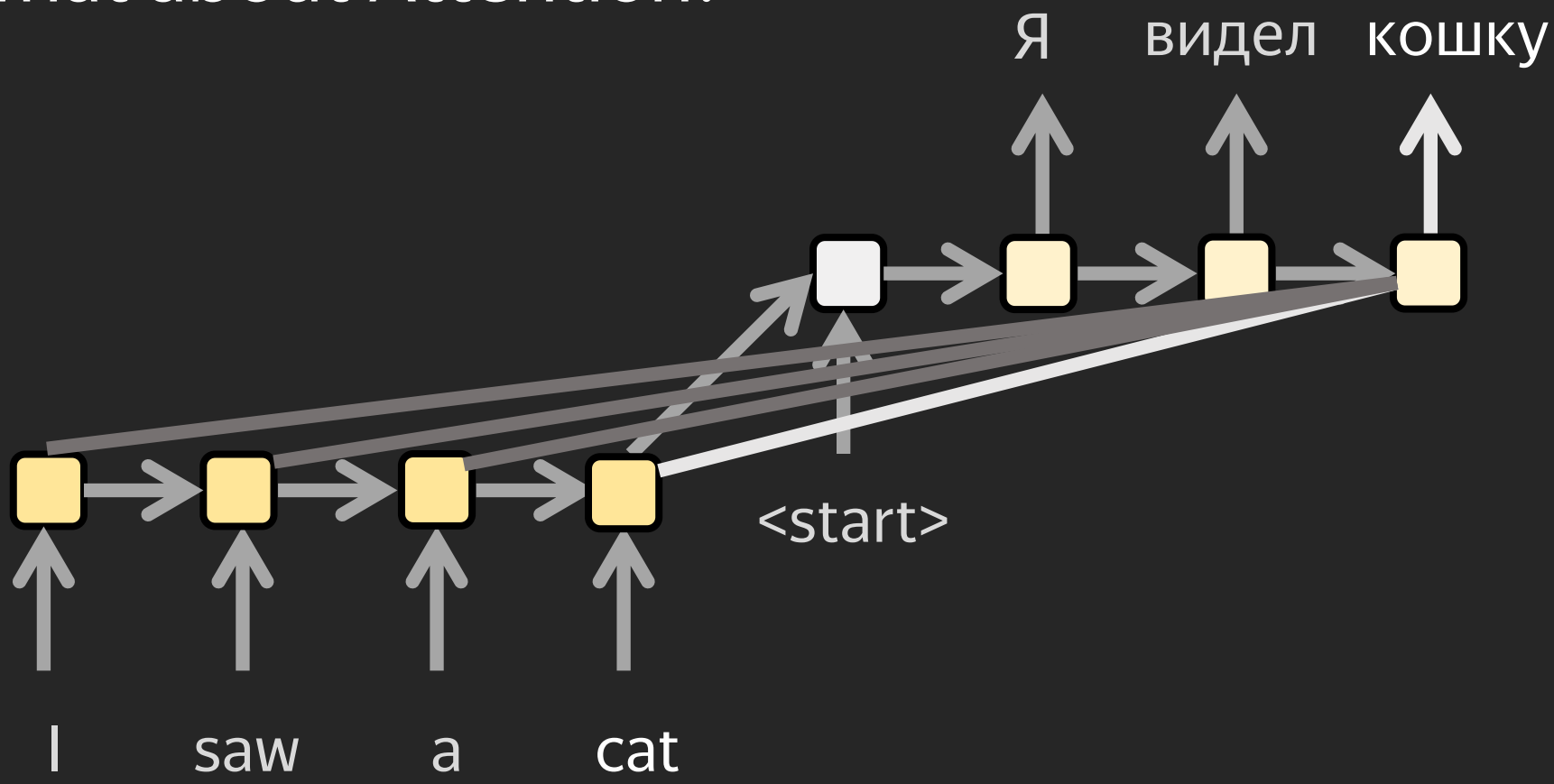
- Much more fluent output

Source: The animal didn't cross the street because it was too tired

PBMT: Животное не пересечь улицу потому, что он слишком устал

NMT: Животное не перешло улицу потому, что оно было слишком уставшим

What about Attention?



Attention vs. Independence

- Attention is a useful inductive bias

Attention vs. Independence

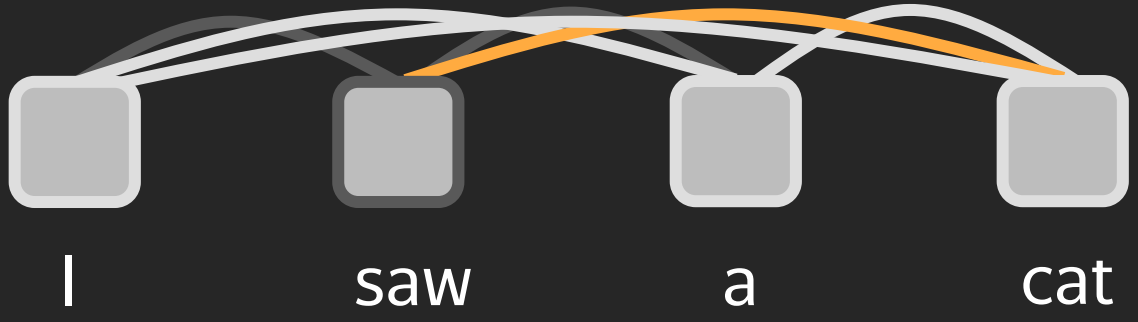
- Attention is a useful inductive bias
- Solves problem of constant capacity in sequence model

Attention vs. Independence

- Attention is a useful inductive bias
- Solves problem of constant capacity in sequence model
- Helps the model decompose the translation process

Multi-Head Self-Attention

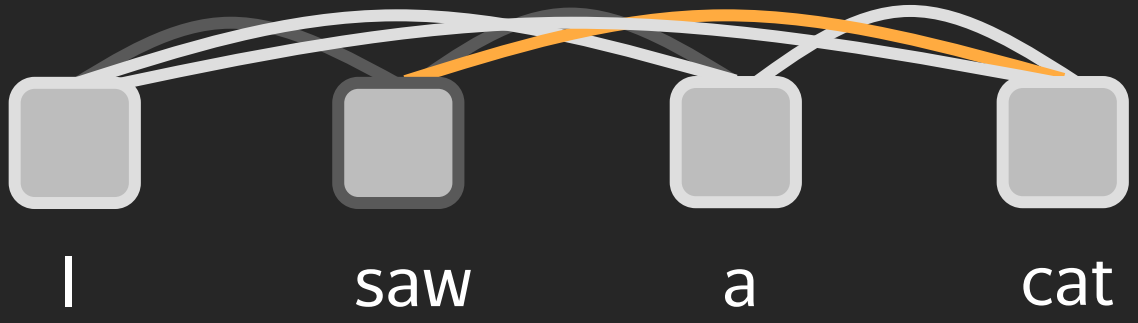
Attention can only really look at **one place** at a time



Multi-Head Self-Attention

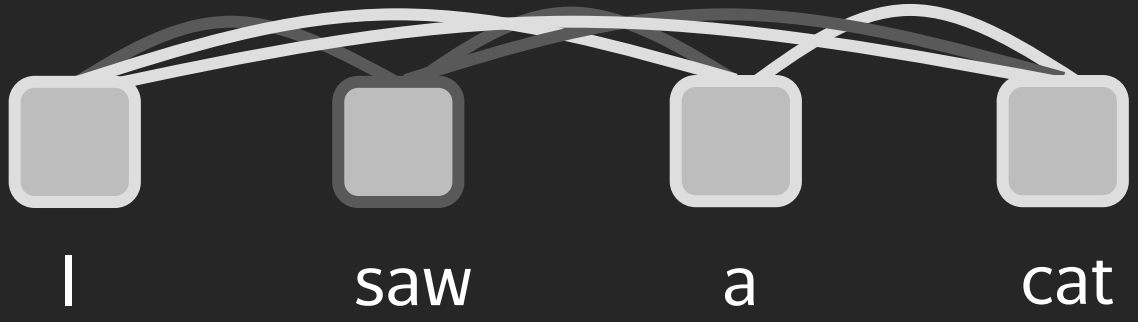
Attention can only really look at **one place** at a time

Attention pulls in information **uniformly** from all dimensions



Multi-Head Self-Attention

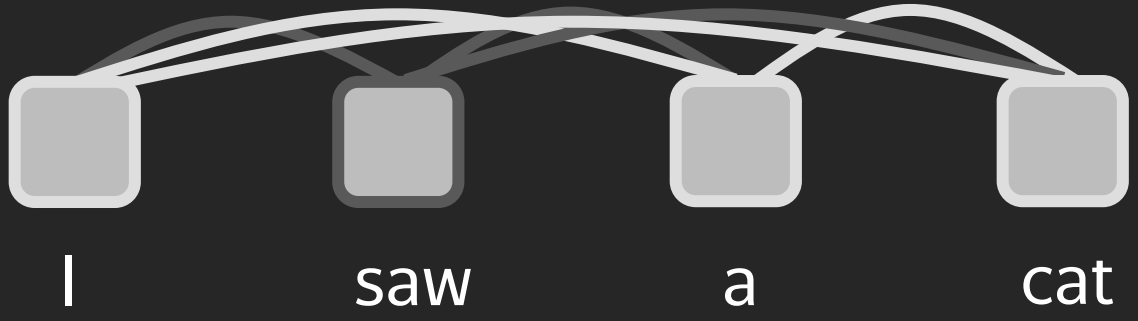
Compute attention in parallel using **different projections**



Multi-Head Self-Attention

Compute attention in parallel using different projections

$$Q_1 = xW_1^Q \qquad K_1 = xW_1^K \qquad V_1 = xW_1^V$$



Multi-Head Self-Attention

Compute attention in parallel using different projections

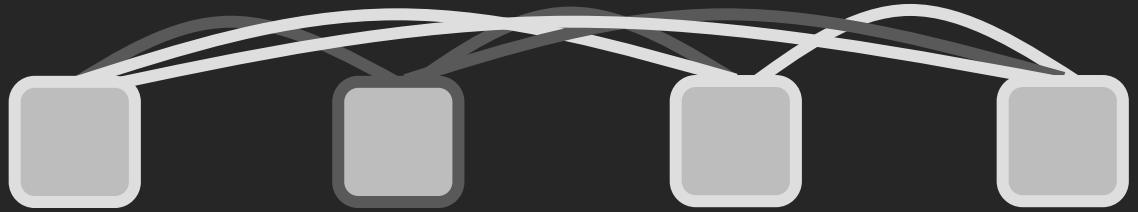
$$Q_1 = xW_1^Q \qquad K_1 = xW_1^K \qquad V_1 = xW_1^V$$

I	saw	a	cat
0.1	0.0	0.2	0.7

Multi-Head Self-Attention

$$Q_1 = xW_1^Q \qquad K_1 = xW_1^K \qquad V_1 = xW_1^V$$

$$Q_2 = xW_2^Q \qquad K_2 = xW_2^K \qquad V_2 = xW_2^V$$

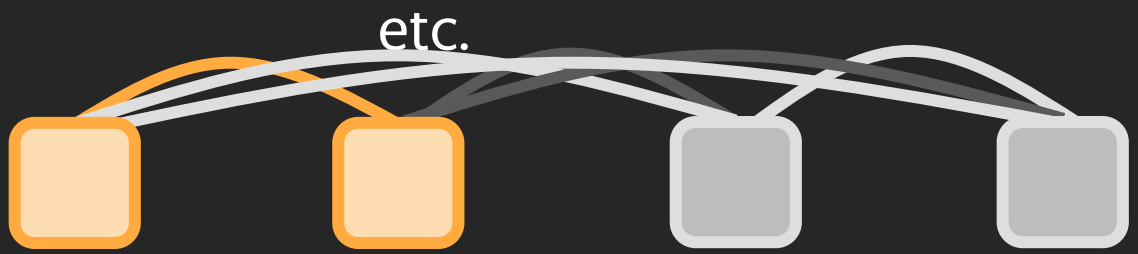


I	saw	a	cat
0.1	0.0	0.2	0.7

Multi-Head Self-Attention

$$Q_1 = xW_1^Q \qquad K_1 = xW_1^K \qquad V_1 = xW_1^V$$

$$Q_2 = xW_2^Q \qquad K_2 = xW_2^K \qquad V_2 = xW_2^V$$



I	saw	a	cat	
0.1	0.0	0.2	0.7	
0.8	0.1	0.1	0.0	etc.

What can multi-head attention do?

- Each transformation can look at **different positions**

What can multi-head attention do?

- Each transformation can look at **different positions**
- Each transformation can focus on **different dimensions**

What does multi-head attention solve?

Она занимается **новым** проектом

- Gender and number agreement
- Case government

Monolingual data in NMT

Monolingual data in NMT

- Not immediately clear how to use it in a direct model of $Pr(e|f)$

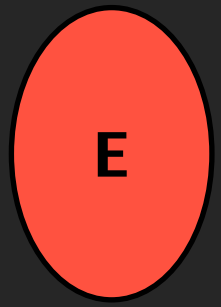
Monolingual data in NMT

- Not immediately clear how to use it in a direct model of $Pr(e|f)$
- Backtranslation most successful approach

Backtranslation in NMT

Backtranslation in NMT

Target language corpus



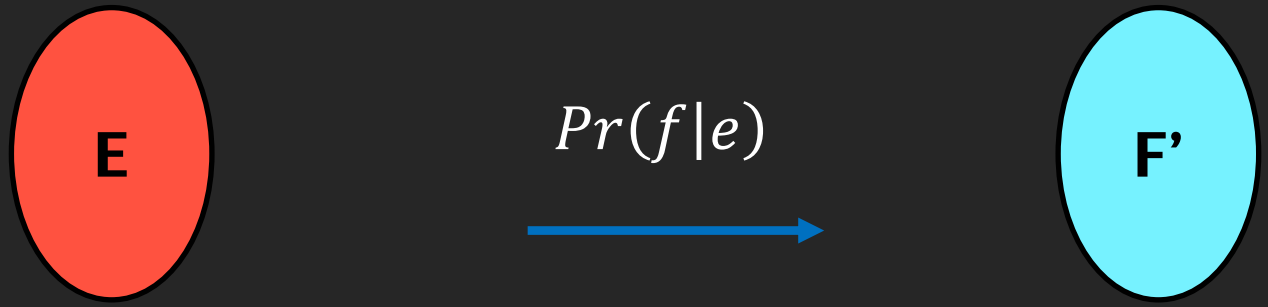
Backtranslation in NMT

Target language corpus

Backtranslation in NMT

Target language corpus

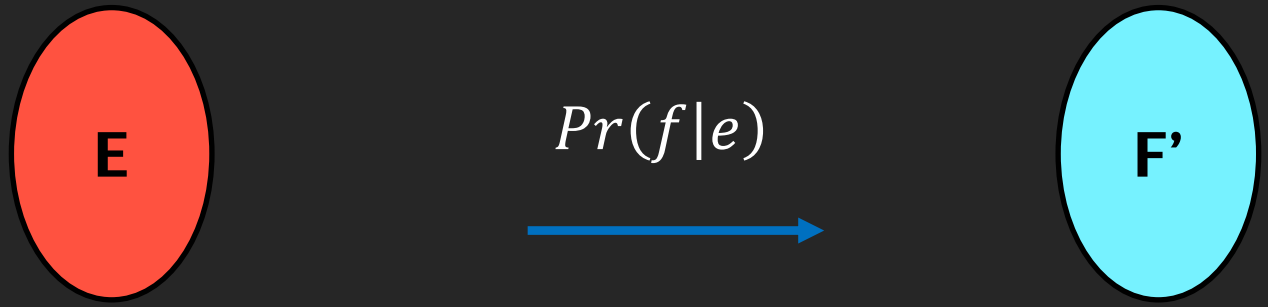
Synthetic source corpus



Backtranslation in NMT

Target language corpus

Synthetic source corpus

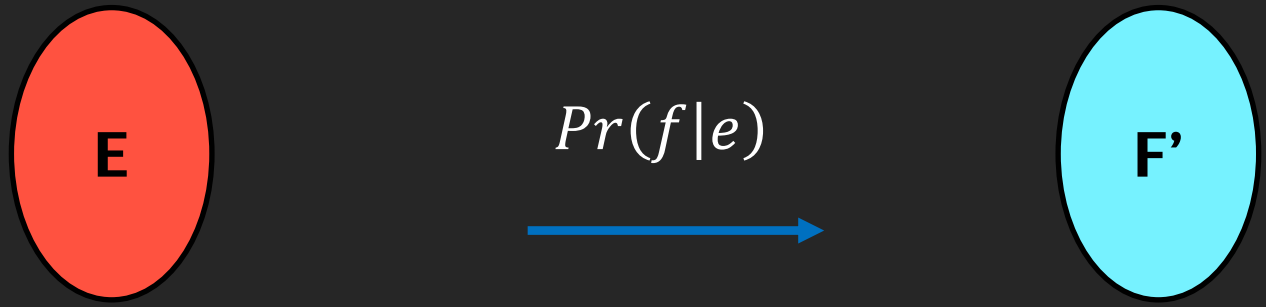


- Train $Pr(e|f)$ on synthetic parallel corpus (E, F')

Backtranslation in NMT

Target language corpus

Synthetic source corpus

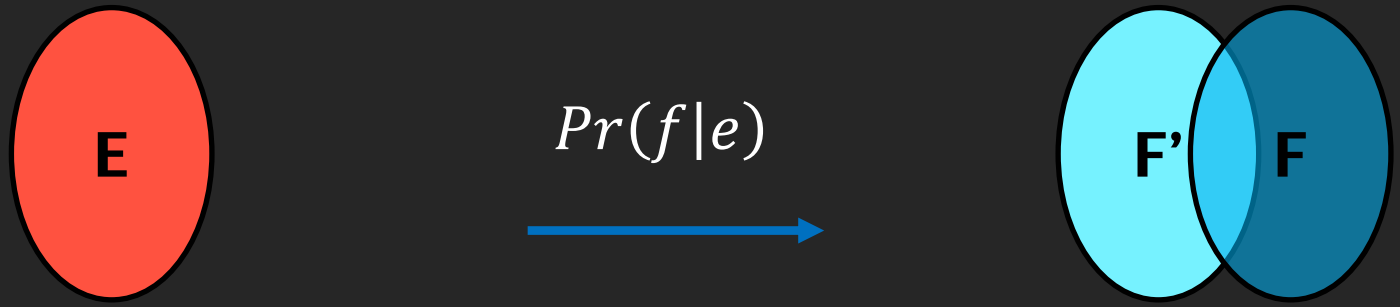


- Train $Pr(e|f)$ on synthetic parallel corpus (E, F')
- Errors in synthetic source corpus F' don’t affect $Pr(e|f)$ for $f \in F$

Backtranslation in NMT

Target language corpus

Synthetic source corpus

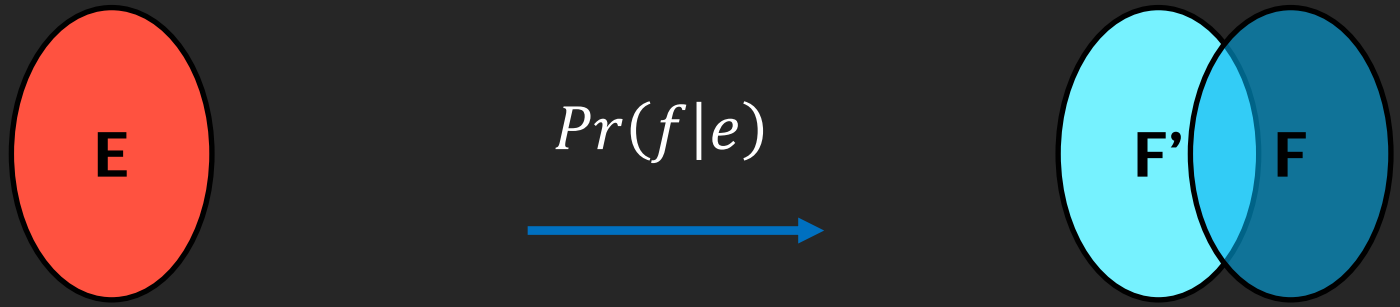


- Train $Pr(e|f)$ on synthetic parallel corpus (E, F')
- Errors in synthetic source corpus F' don't affect $Pr(e|f)$ for $f \in F$

Backtranslation in NMT

Target language corpus

Synthetic source corpus



- Train $Pr(e|f)$ on synthetic parallel corpus (E, F')
- Errors in synthetic source corpus F' don't affect $Pr(e|f)$ for $f \in F$
- Sampling from $Pr(f|e)$ works better than greedy or beam search

Challenges for Neural Machine Translation

Hallucinations

Hallucinations

- Conditional language models can have minds of their own

Hallucinations

- Conditional language models can have minds of their own
- No guarantee that all words are translated

Hallucinations

- Conditional language models can have minds of their own
- No guarantee that all words are translated
- Deep neural models are very good are memorizing random noise

Hallucinations

- Conditional language models can have minds of their own
- No guarantee that all words are translated
- Deep neural models are very good are memorizing random noise
- Increasing NMT beam size beyond ~10 often hurts

Hallucinations - Detached

Source: das kann man nur feststellen , wenn die kontrollen mit einer großen intensität durchgeführt werden .

Correct: this can only be detected if controls undertaken are more rigorous .

Hallucinations - Detached

Source: das kann man nur feststellen , wenn die kontrollen mit einer großen intensität durchgeführt werden .

Correct: this can only be detected if controls undertaken are more rigorous .

Actual: blood alone moves the wheel of history , i say to you and you will understand , it is a privilege to fight .

Hallucinations - Oscillatory

Source: 1995 das produktionsvolumen von 30 millionen pizzen wird erreicht.

Correct: 1995 the production reached 30 million pizzas.

Hallucinations - Oscillatory

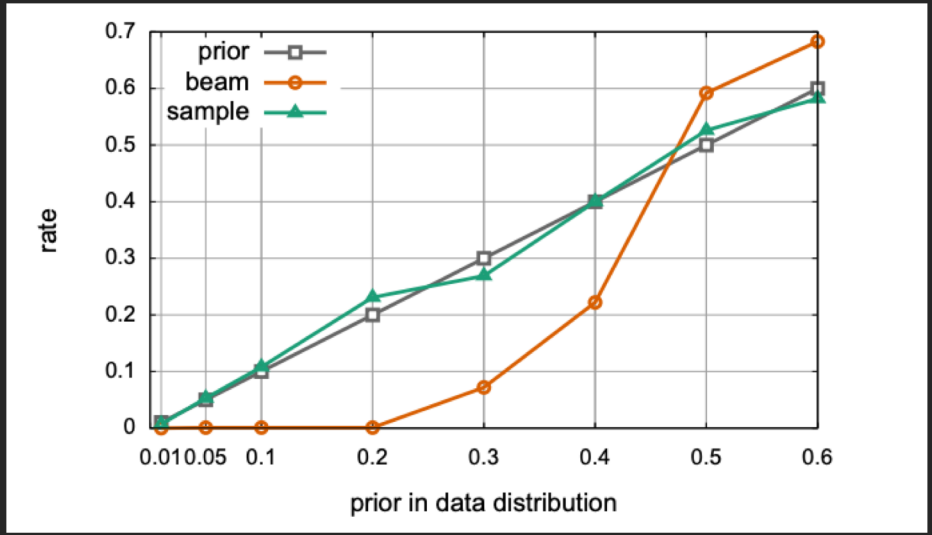
Source: 1995 das produktionsvolumen von 30 millionen pizzen wird erreicht.

Correct: 1995 the production reached 30 million pizzas.

Actual: the us , for example , has been in the past two decades , but has been in the same position as the us , and has been in the united states .

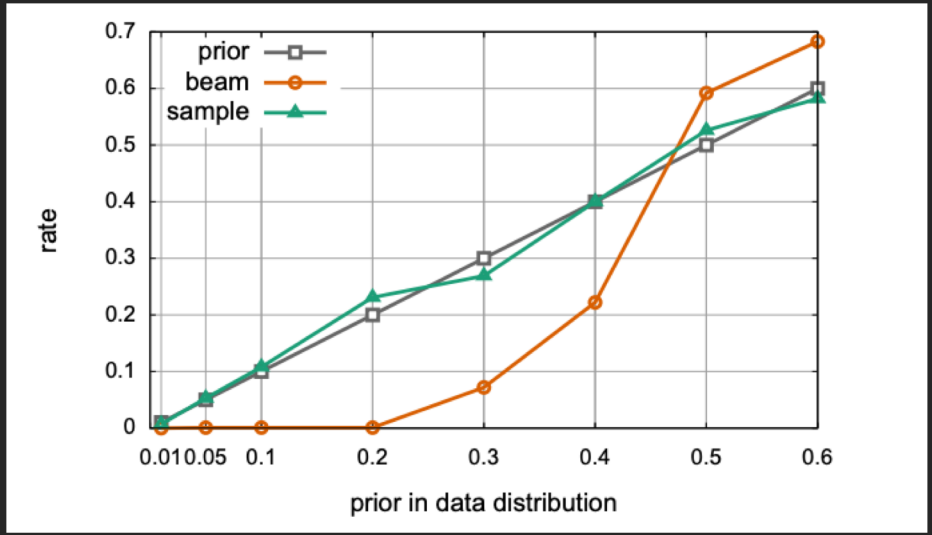
Problems with Beam Search

- Prefers very frequent tokens



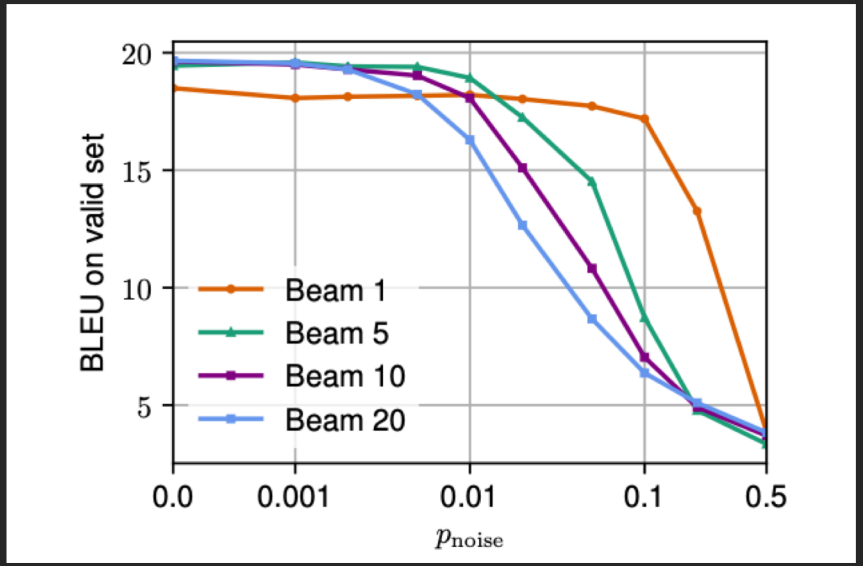
Problems with Beam Search

- Prefers very frequent tokens
- Doesn’t produce diverse hypotheses



Beam size and Noise

- Increasing beam-size hurts even more when training data is noisy (and it is)



Plausible Cause 1: Exposure Bias

Exposure Bias

- Training uses reference as prefix

Exposure Bias

- Training uses reference as prefix
- Inference uses generated result

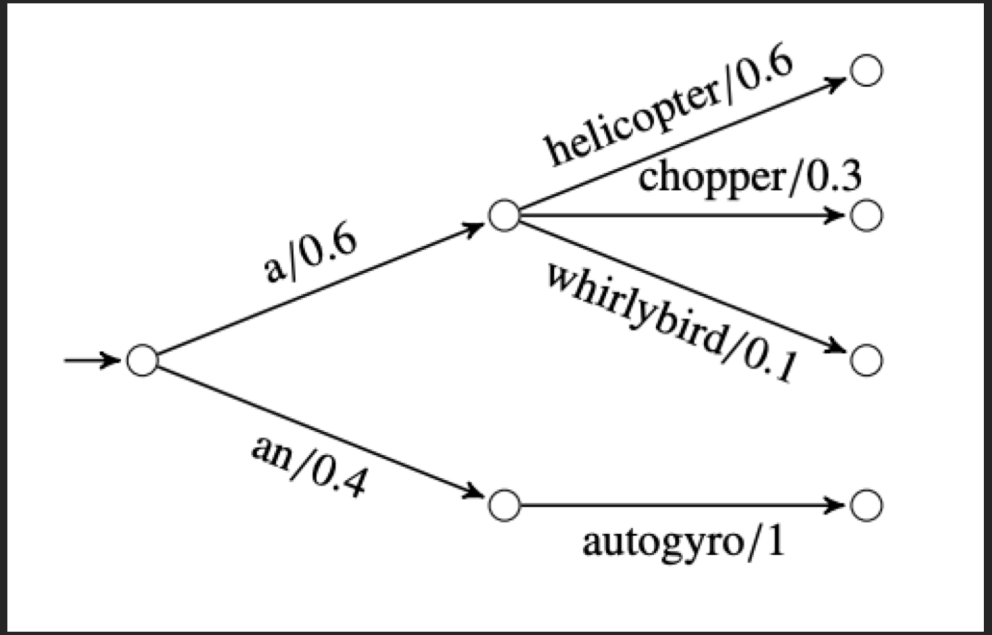
Exposure Bias

- Training uses reference as prefix
- Inference uses generated result
- Model learns to rely too heavily on prefix and then ignore the source

Plausible Cause 2: Label Bias

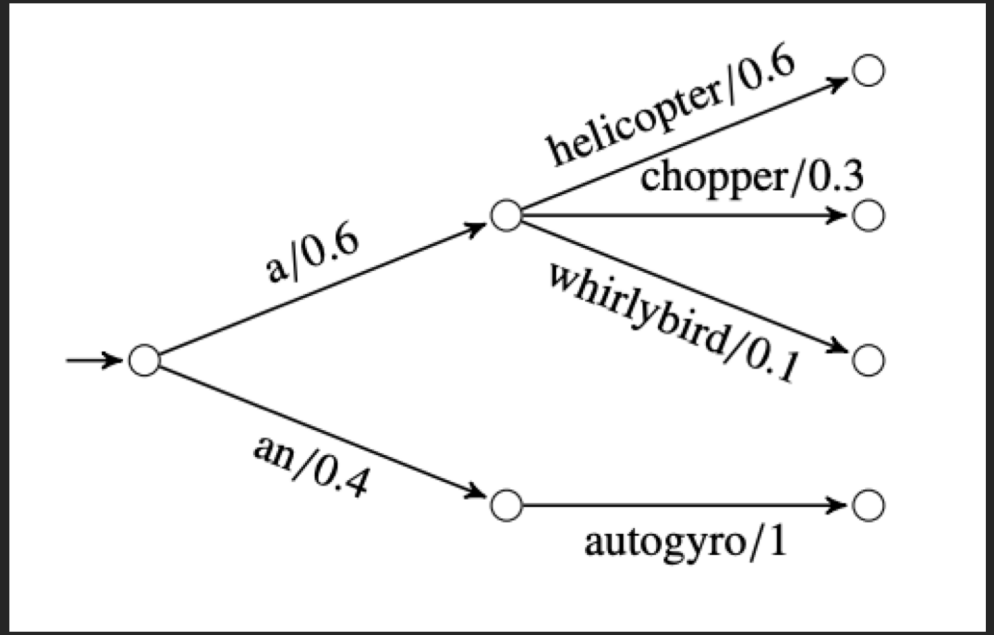
Label Bias

- Example toy word-to-word translation problem



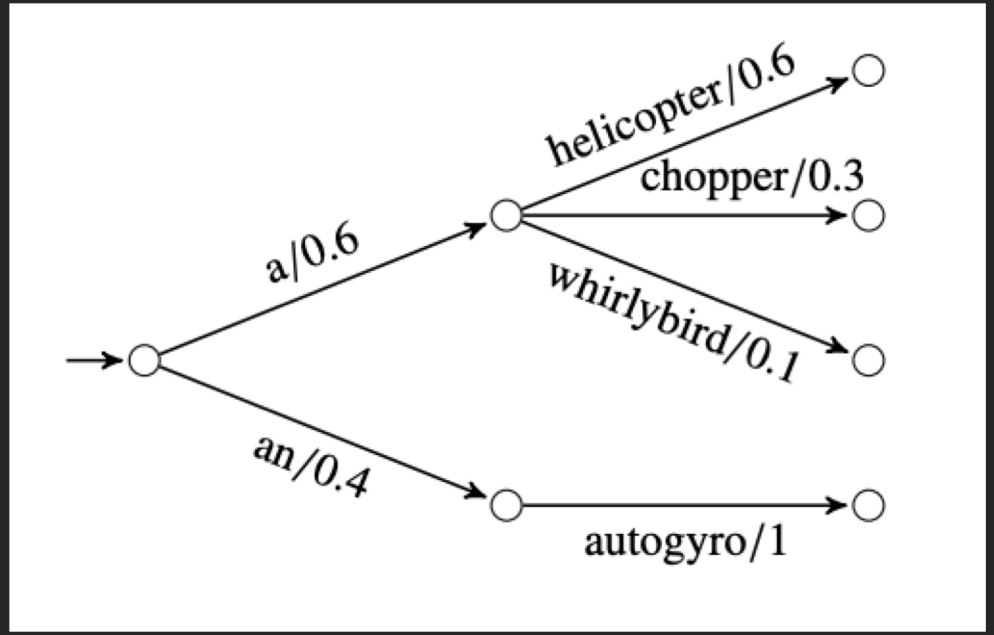
Label Bias

- Example toy word-to-word translation problem
- Locally normalized models can be biased towards states with low-entropy successors



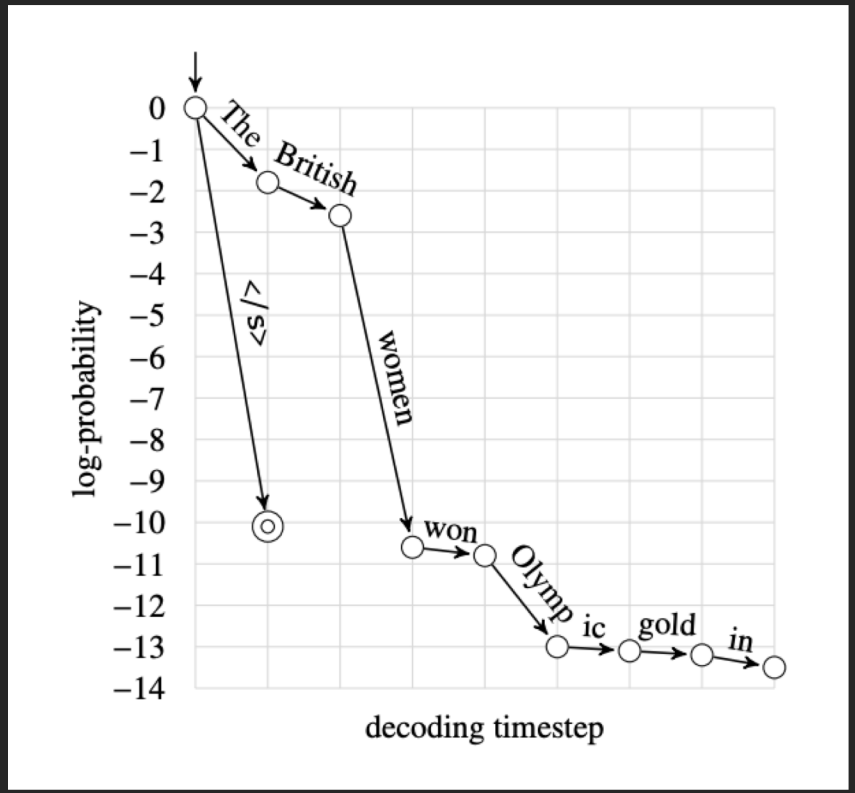
Label Bias

- Example toy word-to-word translation problem
- Locally normalized models can be biased towards states with low-entropy successors
- Current state *explains away* observations



Length Bias (extreme Label Bias)

- Empty hypothesis receives globally best score for >50% of standard test set (WMT15)



Neural Noisy Channel (Yu et al., 2017)

Neural Noisy Channel

- No guarantee in conditional language model $Pr(e|f)$ that all f are accounted for (i.e. *detached hallucinations* are possible)

Neural Noisy Channel

- No guarantee in conditional language model $Pr(e|f)$ that all f are accounted for (i.e. *detached hallucinations* are possible)
- Inverse channel model $Pr(f|e)$ may help solve this problem

Neural Noisy Channel

- No guarantee in conditional language model $Pr(e|f)$ that all f are accounted for (i.e. *detached hallucinations* are possible)
- Inverse channel model $Pr(f|e)$ may help solve this problem

$$e^* = argmax_e \; Pr(e|f)$$
$$= argmax_e \; \underbrace{Pr(e)}_{\text{source}} \; \underbrace{Pr(f|e)}_{\text{channel}}$$

Neural Noisy Channel

- Problems?

$$e^* = argmax_e \; Pr(e|f)$$
$$= argmax_e \; \underbrace{Pr(e)}_{\text{source}} \; \underbrace{Pr(f|e)}_{\text{channel}}$$

Neural Noisy Channel

- Computationally challenging to use $Pr(f|e)$ model at inference time

$$e^* = argmax_e \; Pr(e|f)$$
$$= argmax_e \; \underbrace{Pr(e)}_{\text{source}} \; \underbrace{Pr(f|e)}_{\text{channel}}$$

Neural Noisy Channel

- Computationally challenging to use $Pr(f|e)$ model at inference time
- Sample from proposal distribution then rescore with $Pr(f|e)$

$$e^* = argmax_e \; Pr(e|f)$$
$$= argmax_e \; \underbrace{Pr(e)}_{\text{source}} \; \underbrace{Pr(f|e)}_{\text{channel}}$$

Neural Noisy Channel

- Computationally challenging to use $Pr(f|e)$ model at inference time
- Sample from proposal distribution then rescore with $Pr(f|e)$
- Enables significant improvements with larger beam sizes

$$e^* = argmax_e \; Pr(e|f)$$
$$= argmax_e \; \underbrace{Pr(e)}_{\text{source}} \; \underbrace{Pr(f|e)}_{\text{channel}}$$

Document Context

Document Context

- NMT has gone a long way to approaching the *Bayes error* on the sentence translation problem

Document Context

- NMT has gone a long way to approaching the *Bayes error* on the sentence translation problem
- Humans usually consume and translate text in context

Document Context

- NMT has gone a long way to approaching the *Bayes error* on the sentence translation problem
- Humans usually consume and translate text in context
- Current NMT doesn’t handle context well

Some things we can’t translate without more context

Some things we can’t translate without more context

- Anaphora

“I saw it yesterday”  {он, она, оно, они}

Some things we can’t translate without more context

- Anaphora

  “I saw it yesterday” → {он, она, оно, они}

- Ellipsis

  “I do.” → {?, ?, ?, ...}

Some things we can’t translate without more context

- Anaphora

  “I saw it yesterday” → {он, она, оно, они}

- Ellipsis

  “I do.” → {?, ?, ?, ...}

- Consistency

  “Apple” → {Apple, компания Apple, Яблоко}

Document Level NMT

Document Level NMT

- Parallel document data is not very abundant

Document Level NMT

- Parallel document data is not very abundant
- Solution: source-channel decomposition

Noisy Channel Document NMT

Noisy Channel Document NMT

- Train a source model on monolingual documents

Noisy Channel Document NMT

- Train a source model on monolingual documents
- Train a channel model on parallel sentence corpora

Noisy Channel Document NMT

- Train a source model on monolingual documents
- Train a channel model on parallel sentence corpora

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

Noisy Channel Document NMT

- Train a source model on monolingual documents
- Train a channel model on parallel sentence corpora

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

Noisy Channel Document NMT

- Train a source model on monolingual documents
- Train a channel model on parallel sentence corpora

$$
\begin{aligned}
e^* &= argmax_e \; Pr(e|f) \\
&= argmax_e \; Pr(e) \, Pr(f|e) \\
&\approx argmax_e \; Pr(e) \prod_{i=1}^{I} Pr(f_i|e_i)
\end{aligned}
$$

Noisy Channel Document NMT

- Train a source model on monolingual documents
- Train a channel model on parallel sentence corpora

$$
\begin{aligned}
e^* &= argmax_e \; Pr(e|f) \\
&= argmax_e \; Pr(e) \, Pr(f|e) \\
&\approx argmax_e \; \underbrace{Pr(e)}_{\text{Document-level source model}} \prod_{i=1}^{I} Pr(f_i|e_i)
\end{aligned}
$$

Noisy Channel Document NMT

- Train a source model on monolingual documents
- Train a channel model on parallel sentence corpora

$$
\begin{aligned}
e^* &= argmax_e \; Pr(e|f) \\
&= argmax_e \; Pr(e) \, Pr(f|e) \\
&\approx argmax_e \; \underbrace{Pr(e)}_{\text{Document-level source model}} \; \underbrace{\prod_{i=1}^{I} Pr(f_i|e_i)}_{\text{Sentence-level channel model}}
\end{aligned}
$$

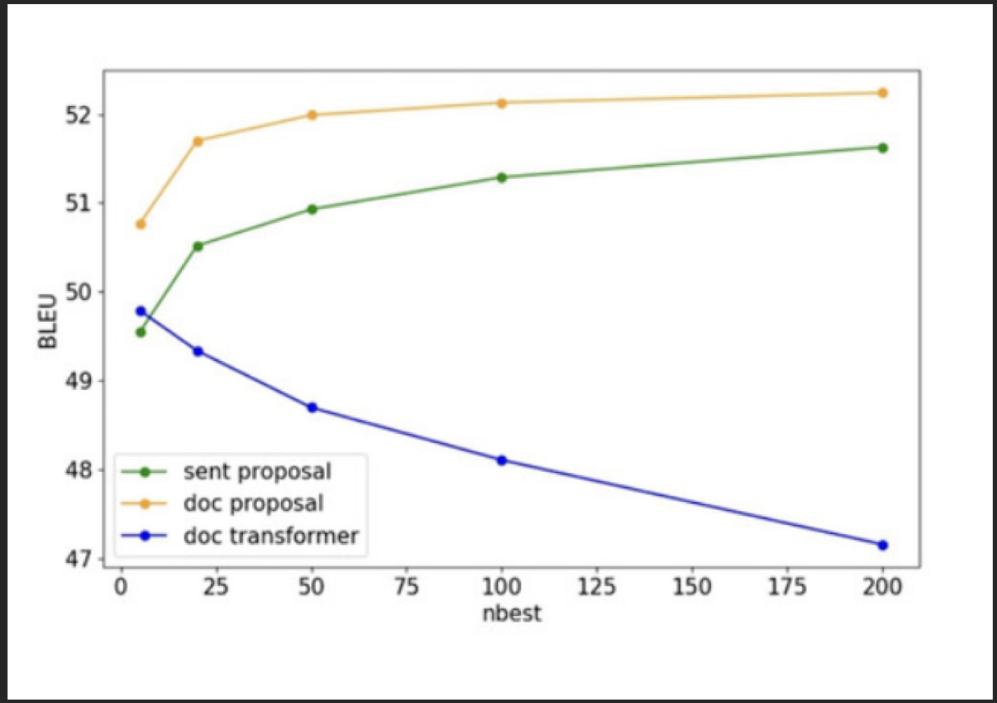
Noisy Channel Document NMT

- Train a source model on monolingual documents
- Train a channel model on parallel sentence corpora
- Independence assumptions in channel okay (remember why?)

$$
\begin{aligned}
e^* &= argmax_e \; Pr(e|f) \\
&= argmax_e \; Pr(e) \, Pr(f|e) \\
&\approx argmax_e \; \underbrace{Pr(e)}_{\text{Document-level source model}} \; \underbrace{\prod_{i=1}^{I} Pr(f_i|e_i)}_{\text{Sentence-level channel model}}
\end{aligned}
$$

Noisy Channel Document NMT

- Generate n-best list with direct model
- Re-rank with noisy channel



Noisy Channel Document NMT

- Increasing beam size no longer hurts