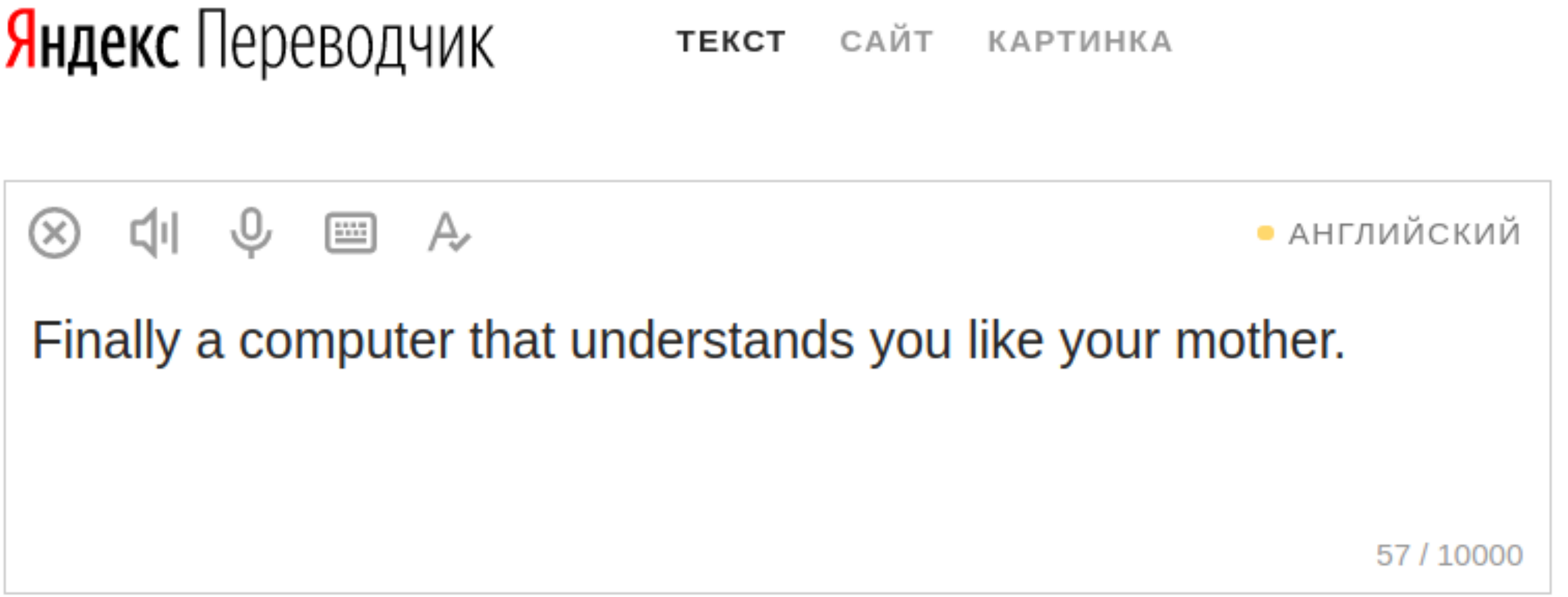
Yandex Translate

Machine Translation

David Talbot

Яндекс

Why Machine Translation is hard


Яндекс Переводчик
ТЕКСТ
САЙТ
КАРТИНКА
АНГЛИЙСКИЙ
Finally a computer that understands you like your mother.
57 / 10000

Яндекс Переводчик

ТЕКСТ САЙТ КАРТИНКА

АНГЛИЙСКИЙ

Finally a computer that understands you like your mother.

57 / 10000

РУССКИЙ

Наконец-то компьютер, который понимает, что ты любишь свою мать.

Перевести в Google Bing

Language Is Ambiguous

Finally a computer that understands you like your mother.

- Наконец-то компьютер, который понимает вас так же хорошо, как ваша мама (понимает вас).

Language Is Ambiguous

Finally a computer that understands you like your mother.

- Наконец-то компьютер, который понимает вас так же хорошо, как ваша мама (понимает вас).
- Наконец-то компьютер, который понимает, что вам нравится ваша мама.

Language Is Ambiguous

Finally a computer that understands you like your mother.

- Наконец-то компьютер, который понимает вас так же, как ваша мама (понимает вас).
- Наконец-то компьютер, который понимает, что вам нравится ваша мама.
- Наконец-то компьютер, который понимает вас так же хорошо, как он понимает вашу маму.

Languages Are Different

A computer that understands you like your mother

Languages Are Different

A computer that understands you like your mother

[your] mother

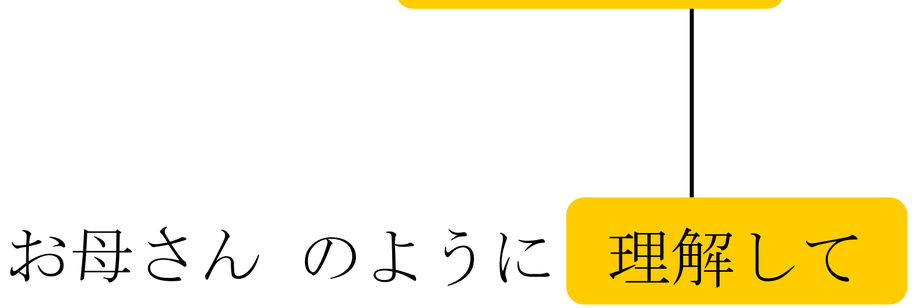
Languages Are Different

A computer that understands you like your mother

[your] mother like

Languages Are Different

A computer that understands you like your mother

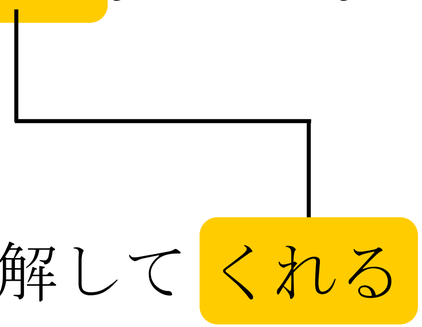


お母さん のように 理解して

[your] mother like understanding

Languages Are Different

A computer that understands you like your mother

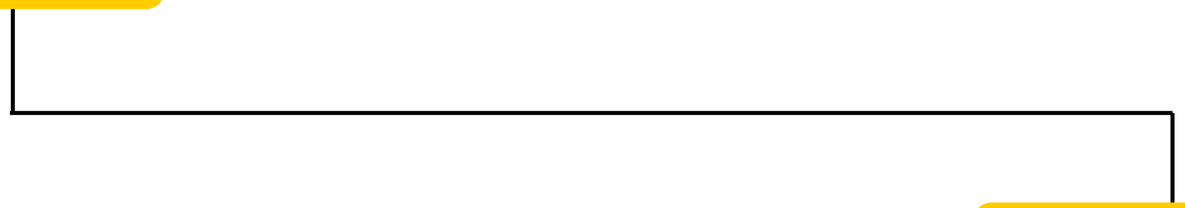


お母さん　のように　理解して くれる

[your] mother like understanding giving

Languages Are Different

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

[your] mother like understanding giving computer

Language Differences

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

Language Differences

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

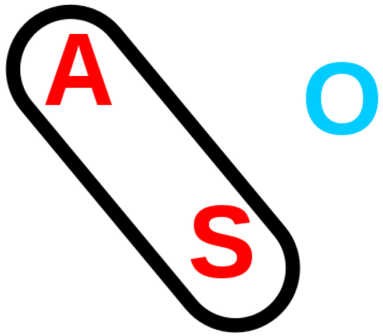
Translate:

Ŋuma banaganyu, yabuŋgu buṛan.

Language Differences

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

Nominative-Accusative



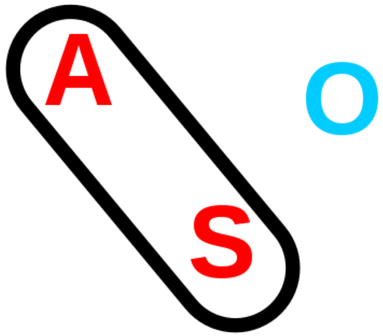
Translate:

Ŋuma banaganyu, yabuŋgu buṛan.
Father returned and saw mother.

Language Differences

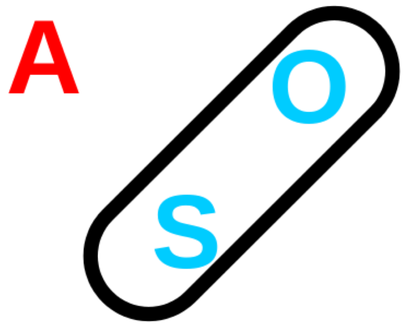
Ŋuma banaganyu. Father returned.
Yabu ŋumaŋgu buṛan. Father saw mother.
Ŋuma yabuŋgu buṛan. Mother saw father.

Nominative-Accusative



Translate:

Ŋuma banaganyu, yabuŋgu buṛan.
~~Father returned and saw mother.~~
Father returned and mother saw [him].



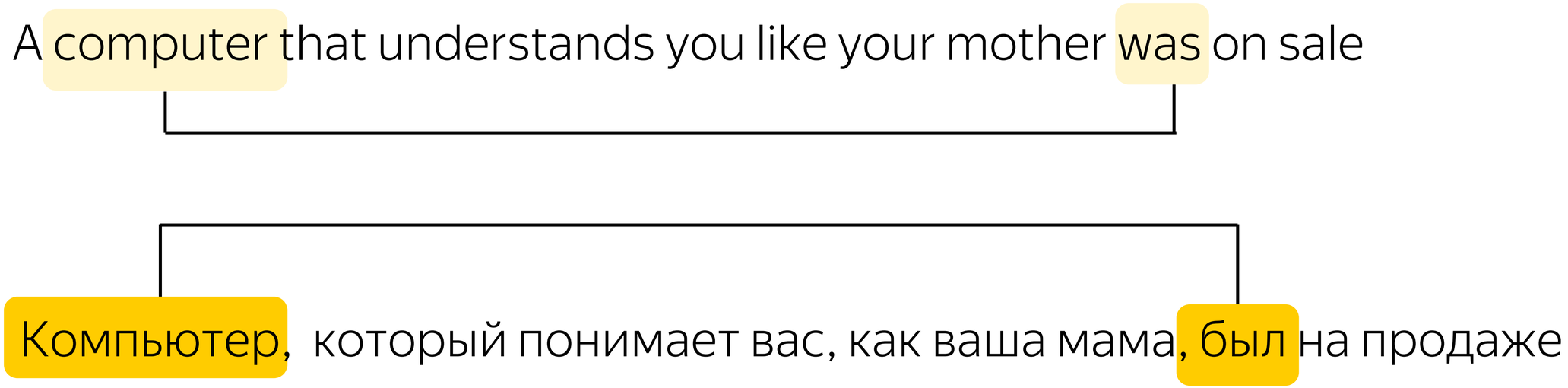
Ergative-Absolute

Long-distance Dependencies

A computer that understands you like your mother was on sale

Компьютер, который понимает вас, как ваша мама, был на продаже

Long-distance Dependencies



Long-distance Dependencies

A computers that understand you like your mother was on sale

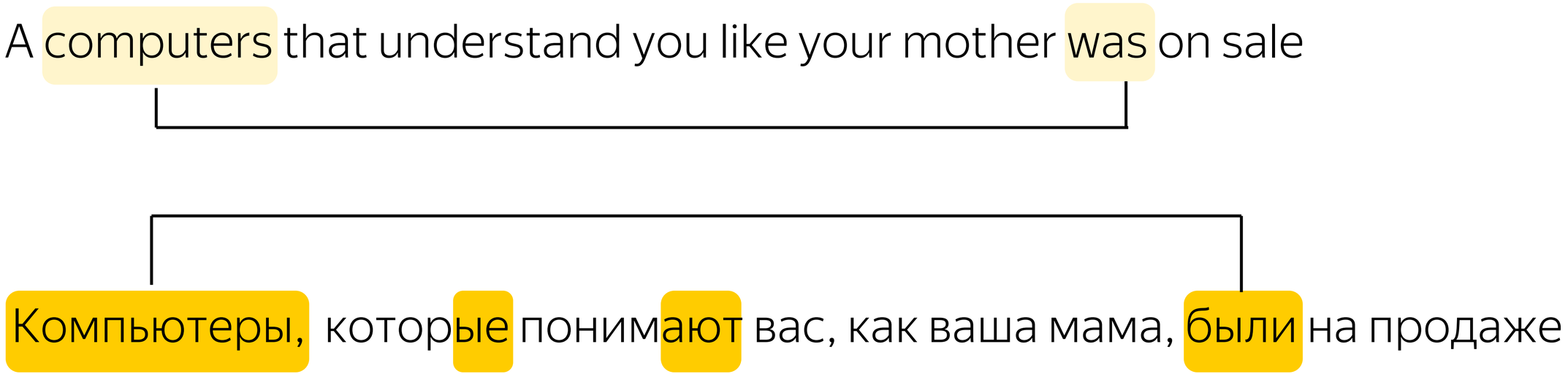
Компьютеры, которые понимают вас, как ваша мама, были на продаже

Long-distance Dependencies

A computers that understand you like your mother was on sale

Компьютеры, которые понимают вас, как ваша мама, были на продаже

Long-distance Dependencies



Discourse Phenomena

- Anaphora

SRC: Does it have a big nose?
TRG: У него | неё большой нос?

Discourse Phenomena

- Anaphora

  SRC: Does it have a big nose?
  TRG: У него | неё большой нос?

- Information Structure

  SRC: Он положил яблоки на стол.
  TRG: He put some | the apples on the table.

Discourse Phenomena

- Anaphora

  SRC: Does it have a big nose?
  TRG: У него | неё большой нос?

- Information Structure

  SRC: Он положил яблоки на стол.
  TRG: He put some | the apples on the table.

- Ellipsis

  SRC: Sure, I did.
  TRG: Конечно, я ему сказал.

Why Evaluating MT is hard

Why Evaluation is Hard for MT

上个星期的战斗至少夺取12个人的生命。

At least 12 people were killed in the battle last week.

Why Evaluation is Hard for MT

上个星期的战斗至少夺取12个人的生命。

At least 12 people were killed in the battle last week.

Last week’s fight took at least 12 lives.

Why Evaluation is Hard for MT

上个星期的战斗至少夺取12个人的生命。

At least 12 people were killed in the battle last week.
Last week's fight took at least 12 lives.
The fighting last week killed at least 12.

Why Evaluation is Hard for MT

上个星期的战斗至少夺取12个人的生命。

At least 12 people were killed in the battle last week.
Last week's fight took at least 12 lives.
The fighting last week killed at least 12.
The battle of last week killed at least 12 persons.

Why Evaluation is Hard for MT

上个星期的战斗至少夺取12个人的生命。

At least 12 people were killed in the battle last week.
Last week's fight took at least 12 lives.
The fighting last week killed at least 12.
The battle of last week killed at least 12 persons.
At least 12 people lost their lives in last week's fighting.
At least 12 persons died in the fighting last week.
At least 12 died in the battle last week.
At least 12 people were killed in the fighting last week.
During last week's fighting, at least 12 people died.
Last week at least twelve people died in the fighting.
Last week's fighting took the lives of twelve people

Human Evaluation

- Fluency: Does the translation sounds good?
- Adequacy: Does the translation preserve the information?

They hated the film	Adequacy	Fluency
Они очень понравился фильм		
Они очень не понравился фильм		
Им очень понравился фильм		
Им очень не понравился фильм		

Human Evaluation

- Fluency: Does the translation sounds good?
- Adequacy: Does the translation preserve the information?

They hated the film	Adequacy	Fluency
Они очень понравился фильм	N	N
Они очень не понравился фильм	Y	N
Им очень понравился фильм	N	Y
Им очень не понравился фильм	Y	Y

Types of Human Evaluation: Pairwise

Source:	Your browser is not supported.
Translation 1.	Ваш браузер не поддерживается.
Translation 2.	Ваш браузер не поддерживаемый.

Choices: (1) One is better (2) Two is better (3) Translations are same

Types of Human Evaluation: Direct Assessment

Reference	Ваш браузер не поддерживается.
Translation 1.	Ваш браузер не поддерживается.
Translation 2.	Ваш браузер не поддерживаемый.

Choices: (1) One is better (2) Two is better (3) Translations are same

Difficulties of Human Evaluation

- Pairwise comparisons requires bilingual raters
- Direct assessment requires references
- Humans often biased towards fluency
- Agreement can be low
- Fraud detection can be hard

BLEU

- N-gram overlap between candidates and reference translations (clipped)
- Compute precision for n-grams of length 1 to 4
- Add brevity penalty if too short
- Compute over the whole corpus (usually)

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right) \left(\prod_{i=1}^{4} \textit{precision}_i\right)^{\frac{1}{4}}$$

BLEU Example

System: I like green apples more than yellow ones.

Reference 1: I prefer green apples to yellow apples.
Reference 2: I would rather have green apples to yellow ones.

Matches
1-grams: {I, green, apples, ones, yellow}
2-grams: {green apples, yellow ones}

BLEU Example

Iran has already stated that Kharazi's statements to the conference because of the Jordanian King Abdullah II in which he stood accused Iran of interfering in Iraqi affairs.
n-gram matches: 27 unigrams, 20 bigrams, 15 trigrams, and ten 4-grams
human scores: Adequacy:3,2 Fluency:3,2
Iran already announced that Kharrazi will not attend the conference because of the statements made by the Jordanian Monarch Abdullah II who has accused Iran of interfering in Iraqi affairs.
n-gram matches: 24 unigrams, 19 bigrams, 15 trigrams, and 12 4-grams
human scores: Adequacy:5,4 Fluency:5,4

Building Better Automatic Metrics

- System level correlation with human judgements

Building Better Automatic Metrics

- System level correlation with human judgements
- Sentence level correlation with human judgements

Building Better Automatic Metrics

- System level correlation with human judgements
- Sentence level correlation with human judgements
- BLEU was first proposed in 2002

Building Better Automatic Metrics

- System level correlation with human judgements
- Sentence level correlation with human judgements
- BLEU was first proposed in 2002
- BLEU was first shown to be broken around 2003

Building Better Automatic Metrics

- System level correlation with human judgements
- Sentence level correlation with human judgements
- BLEU was first proposed in 2002
- BLEU was first shown to be broken around 2003
- WMT Metrics Track has been running for over 5 years

Building Better Automatic Metrics

- System level correlation with human judgements
- Sentence level correlation with human judgements
- BLEU was first proposed in 2002
- BLEU was first shown to be broken around 2003
- WMT Metrics Track has been running for over 5 years
- BLEU is still the dominant metric in the field