Машинный перевод

Сергей Губанов
Яндекс
esgv@yandex-team.ru

9 февраля 2017 г.

План

История

Задача

Качество

Треугольник

Структура курса

План

История

Задача

Качество

Треугольник

Структура курса

"Translation"memorandum (1949)

Warren Weaver

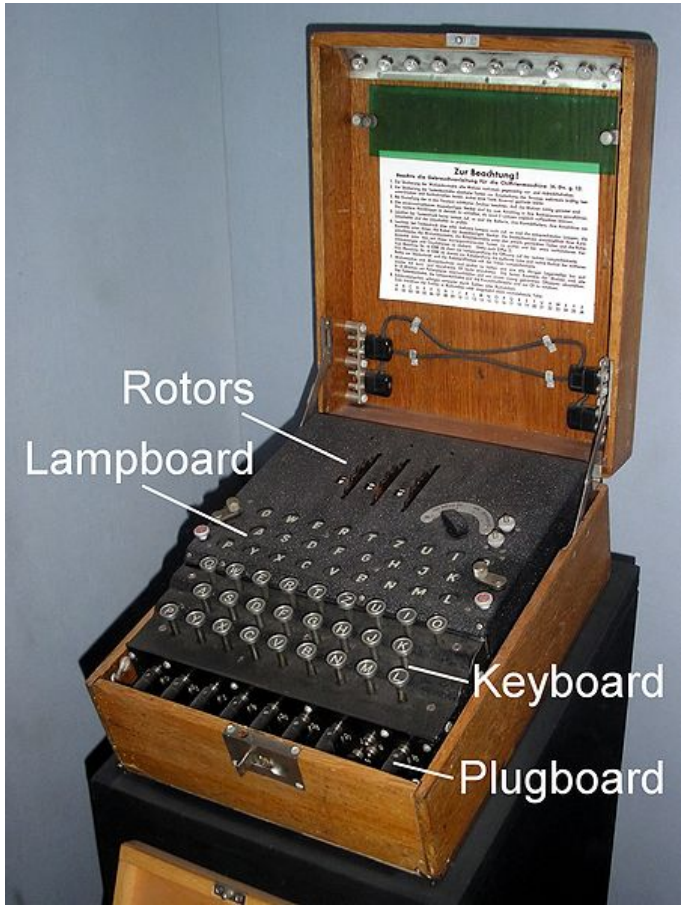
http://www.mt-archive.info/Weaver-1949.pdf

Исторический контекст: вторая мировая, успехи в криптоанализе, теория информации, компьютеры.

- Использовать контекст в предложении.
- Использовать компьютеры.
- Использовать методы криптоанализа.
- Использовать связи между языками.

Enigma & bombe

Bombe



Enigma

"Translation"memorandum (1949)

Warren Weaver

http://www.mt-archive.info/Weaver-1949.pdf

- Использовать контекст в предложении.
- Использовать компьютеры.
- Использовать методы криптоанализа.
- Использовать связи между языками.

"Translation"memorandum (1949)

Warren Weaver

http://www.mt-archive.info/Weaver-1949.pdf

- Использовать контекст в предложении.
- Использовать компьютеры.
- Использовать методы криптоанализа.
- Использовать связи между языками.

Контекст в предложении

"bat"

- I want a new baseball bat.
- I saw a flying vampire bat.

Контекст в предложении

"bat"

- I want a new baseball bat.
- I saw a flying vampire bat.

Translation memorandum (1949)

«This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.»

Translation memorandum (1949)

«This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.»

«A most serious problem, for UNESCO and for the constructive and peaceful future of the planet, is the problem of translation, as it unavoidably affects the communication between peoples.»

Georgetown experiment (1954)

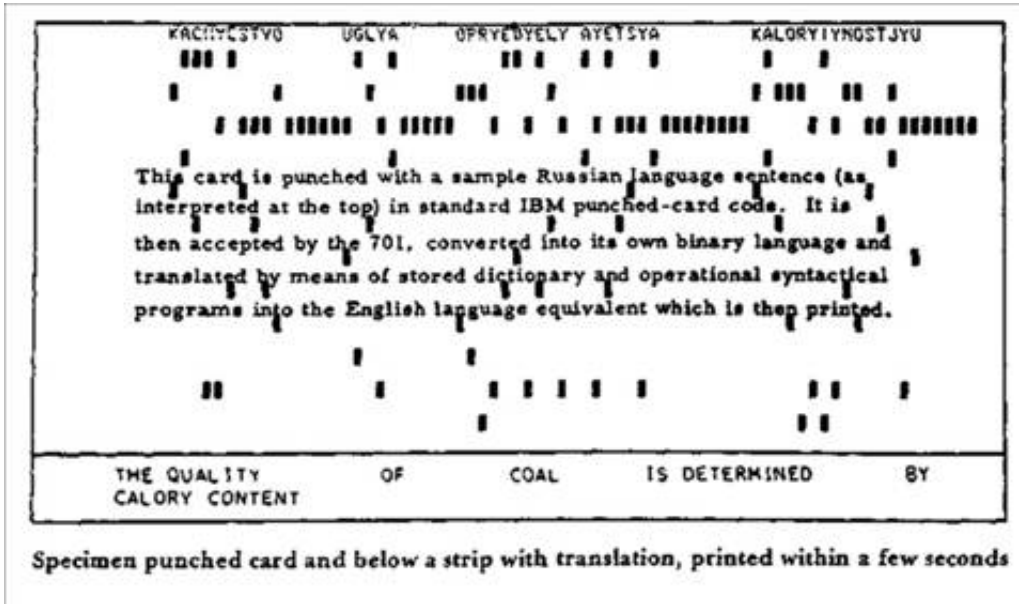
Georgetown university & IBM

http://www.hutchinsweb.me.uk/GU-IBM-2005.pdf

Контекст: начало холодной войны

- Публичная демонстрация в штаб-квартире IBM.
- Советский (т.е. русский) в английский.
- 250 слов в словаре, 6 правил.
- Переведено более 60 предложений (органическая химия, общая тематика).

Georgetown experiment (1954)



Specimen punched card and below a strip with translation, printed within a few seconds

Mi pyeryedayem mislyi posryedstvom ryechyi.
We transmit thoughts by means of speech.

Voyenniy sud pryigovoryil syerzhanta k lyishyenyiyu grazhdanskyix prav
A military court sentenced a sergeant to deprival of civil rights

Reception

«Electronic brain translates Russian»

Reception

«Electronic brain translates Russian»

Организаторы заявляли, что проблема машинного перевода будет полностью решена в следующие 3-5 лет.

Спойлер

(этого не произошло)

ALPAC report (1966)

http://www.hutchinsweb.me.uk/ALPAC-1996.pdf

ALPAC report (1966)

http://www.hutchinsweb.me.uk/ALPAC-1996.pdf

- Пост-редактировать машинный перевод – дольше, чем просто переводить. Придумайте, как ускорить работу человека-переводчика.

ALPAC report (1966)

http://www.hutchinsweb.me.uk/ALPAC-1996.pdf

- Пост-редактировать машинный перевод – дольше, чем просто переводить. Придумайте, как ускорить работу человека-переводчика.
- Следите, используются ли потом переводы, чтобы не переводить тексты впустую.

ALPAC report (1966)

http://www.hutchinsweb.me.uk/ALPAC-1996.pdf

- Пост-редактировать машинный перевод – дольше, чем просто переводить. Придумайте, как ускорить работу человека-переводчика.
- Следите, используются ли потом переводы, чтобы не переводить тексты впустую.
- «There is no emergency in the field of translation. The problem is not to meet some nonexistent need through nonexistent machine translation. There are, however, several crucial problems of translation. These are quality, speed, and cost.»

AI winter

https://en.wikipedia.org/wiki/AI_winter

- 1966: the failure of machine translation,
- 1970: the abandonment of connectionism,
- 1971–75: DARPA's frustration with the Speech Understanding Research program at Carnegie Mellon University,
- 1973: the large decrease in AI research in the United Kingdom in response to the Lighthill report,
- 1973–74: DARPA's cutbacks to academic AI research in general,
- 1987: the collapse of the Lisp machine market,
- 1988: the cancellation of new spending on AI by the Strategic Computing Initiative,
- 1993: expert systems slowly reaching the bottom, and
- 1990s: the quiet disappearance of the fifth-generation computer project's original goals.

late 1960s – mid 1980s

- SYSTRAN, Logos, METEO.
- Микрокомпьютеры.
- Нужно много языков: глобализация, инструкции, документация.

A statistical approach to MT (1988)

Brown, Peter F., et al.
"The mathematics of statistical machine translation: Parameter estimation."
Computational linguistics 19.2 (1993): 263-311.

- $P(e|f) \propto P(f|e)P(e)$, noisy channel model
- IBM Model 1-5.

http://cs.jhu.edu/~post/bitext/

A statistical approach to MT (1988)

COLING review:

The validity of statistical information theoretic approach to machine translation has indeed been recognized as the authors mentioned by Weaver as early as 1949, and was universally recognized as mistaken by 1950. The crude force of computers is not science; the paper is simply beyond the scope of COLING.

A statistical approach to MT (1988)

COLING review:

The validity of statistical information theoretic approach to machine translation has indeed been recognized as the authors mentioned by Weaver as early as 1949, and was universally recognized as mistaken by 1950. **The crude force of computers is not science;** the paper is simply beyond the scope of COLING.

"Translation"memorandum (1949)

Warren Weaver

- Использовать контекст в предложении.
- Использовать компьютеры.
- Использовать методы криптоанализа.
- Использовать связи между языками.

Великое пробуждение искусственного интеллекта (2012 —)

- ImageNet
- Self-driving cars
- Speech recognition
- Neural machine translation
- etc.

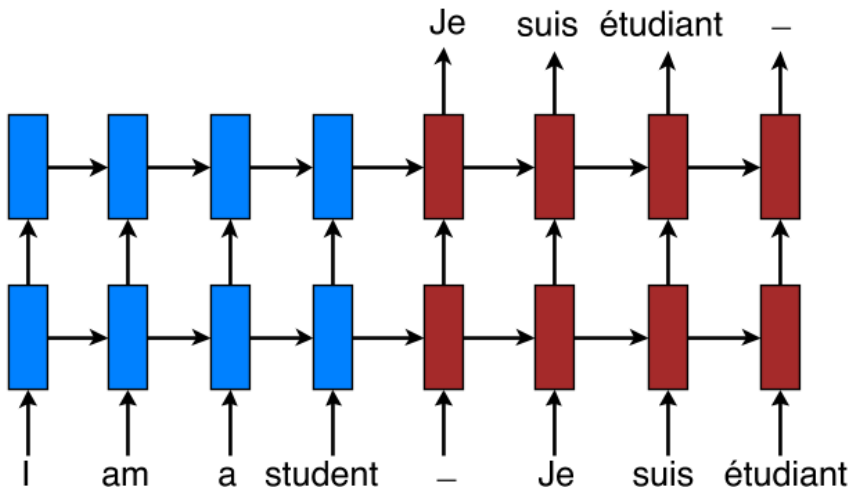
Великое пробуждение искусственного интеллекта (2012 —)

- ImageNet
- Self-driving cars
- Speech recognition
- Neural machine translation
- etc.

Нейросети + **GPU** + Data = ♡

Neural machine translation (2014)

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le.
"**Sequence to sequence learning with neural networks.**"
Advances in neural information processing systems. 2014.



Neural machine translation (2016)

Wu, Yonghui, et al.
"Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation."
arXiv preprint arXiv:1609.08144 (2016).

Neural machine translation (2016)

Wu, Yonghui, et al.
"Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation."
arXiv preprint arXiv:1609.08144 (2016).

Crego, Josep, et al.
"SYSTRAN's Pure Neural Machine Translation Systems".
arXiv preprint arXiv:1610.05540 (2016).

Настоящее время

Покинув кафе, шестеро мужчин, одетых в одинаковые чёрные костюмы, совершают вооружённое ограбление ювелирной лавки среди бела дня. Но на месте преступления их уже поджидает полиция. Тщательно просчитанный план срывается, и грабители, потеряв в перестрелках двух своих подельников (мистера Синего и мистера Коричневого), незамедлительно скрываются с места преступления.

After leaving the cafe, six men, dressed in identical black suits, commit an armed robbery of a jewelry shop in broad daylight. But at the scene they were waiting for the police. A carefully calculated plan breaks down, and the robbers, having lost in the shootings of two of his accomplices (Mr. Blue and Mr. Brown), immediately disappear from the scene.

Word lens

By Quest Visual, Inc. ZTebaykina at en.wikipedia - Word Lens demo. Transferred from en.wikipedia by Ronhjones, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=18310222

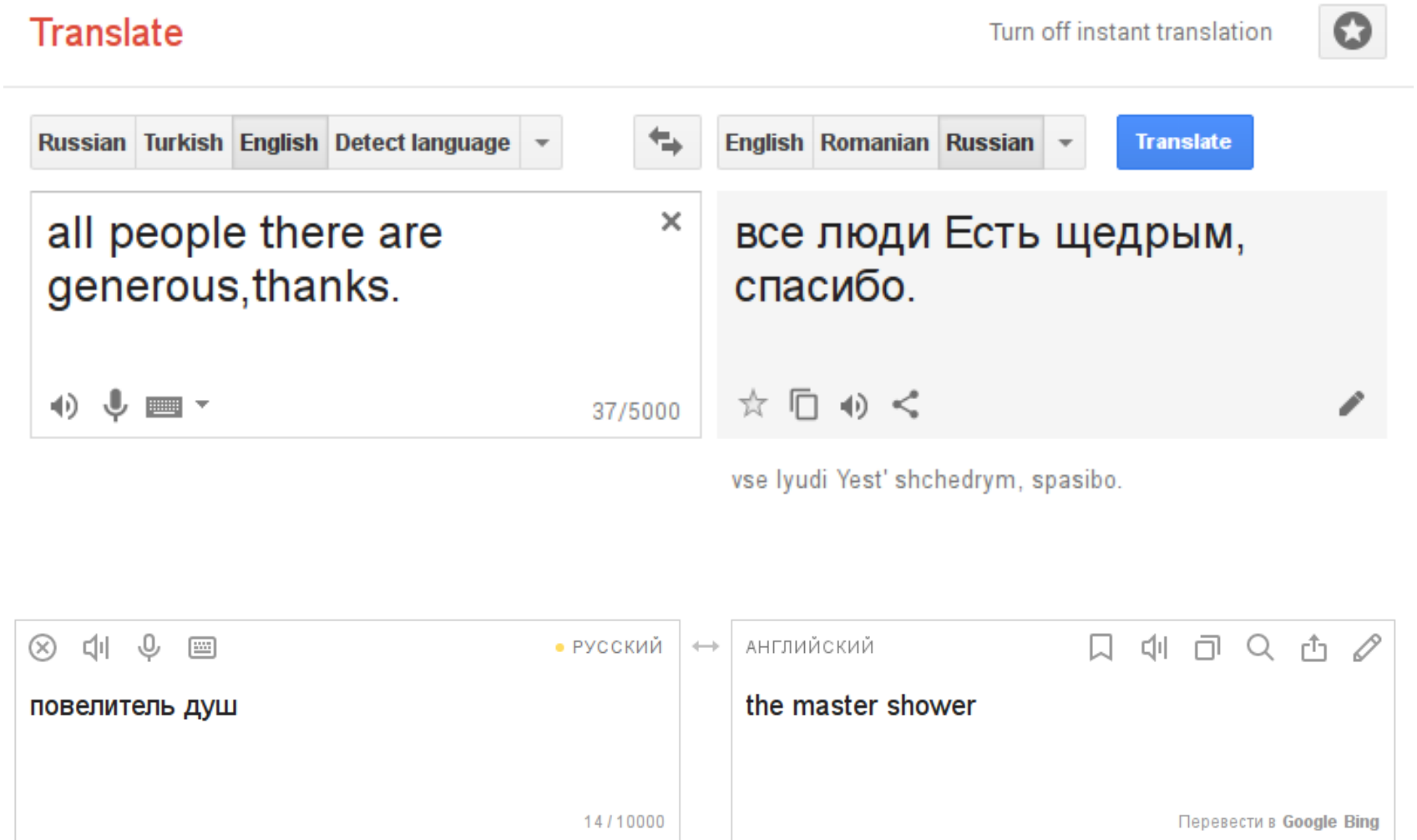
Speech-to-speech

Speech-to-speech translation by MSR.

https://www.youtube.com/watch?v=Nu-nlQqFCKg

Skype translator

Bridging the gap



План

История

Задача

Качество

Треугольник

Структура курса

Порядок слов

Порядок слов

«İstikbal göklerdedir.»

— Mustafa Kemal Atatürk

The future is in the skies.

Порядок слов

«İstikbal gökler(de)(dir).»

— Mustafa Kemal Atatürk

The future is in the skies.

The future the skies in is.

Порядок слов

«İstikbal gökler(de)(dir).»

— Mustafa Kemal Atatürk

The future is in the skies.

The future the skies in [is].

Порядок слов

«İstikbal gökler(de)(dir).»

— Mustafa Kemal Atatürk

The future is in the skies.

The future [the skies in] [is].

Порядок слов

«İstikbal gökler(de)(dir).»

— Mustafa Kemal Atatürk

The future is in the skies.

[The future] [the skies in] [is].

Разная синтаксическая структура

SVO vs SOV

John read the letter

John ga tegami o yon-da

Омонимия

He deposited money in a bank account with a high interest rate.

Sitting on the bank of the Mississippi, a passing ship piqued his interest

Согласование

Vladimir appears for work late in the morning

Владимир является на работу поздно утром

Идиомы

It's raining cats and dogs today.

Сегодня дождь льёт как из ведра.

Machine translation is a piece of cake.

Машинный перевод – несложная задача.

Синтаксическая неоднозначность

I saw a man with a telescope.

- I saw [a man] [with a telescope].
- I saw [a man [with a telescope]].

Компьютерная *лингвистика*

Много разных языков (100+), похожи друг на друга.

- Гласные и согласные

Компьютерная *лингвистика*

Много разных языков (100+), похожи друг на друга.

- Гласные и согласные
- Порядок слов: SOV, SVO, etc.
  - John read the letter
  - John ga tegami o yon-da

Компьютерная *лингвистика*

Много разных языков (100+), похожи друг на друга.

- Гласные и согласные
- Порядок слов: SOV, SVO, etc.
  - John read the letter
  - John ga tegami o yon-da
- Выражение отрицания
  - I do **not** see the moon.
  - Je **ne** vois **pas** la lune.

Компьютерная *лингвистика*

Много разных языков (100+), похожи друг на друга.

- Гласные и согласные
- Порядок слов: SOV, SVO, etc.
  - John read the letter
  - John ga tegami o yon-da

  http://wals.info/chapter/81
- Выражение отрицания
  - I do **not** see the moon.
  - Je **ne** vois **pas** la lune.

  http://wals.info/chapter/112

http://wals.info/

"Translation"memorandum (1949)

Warren Weaver

- Использовать контекст в предложении.
- Использовать компьютеры.
- Использовать методы криптоанализа.
- Использовать связь между языками.

Компьютерная лингвистика

Знаний про языки не требуется.

Декодирование

"Декодирование"= процесс перевода.

Декодирование

"Декодирование"= процесс перевода.

«This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.»

Английский со словарем

Пусть у нас есть словарь (т.е. перевод для каждого слова).

Дороги *строятся* *из* *бетона*

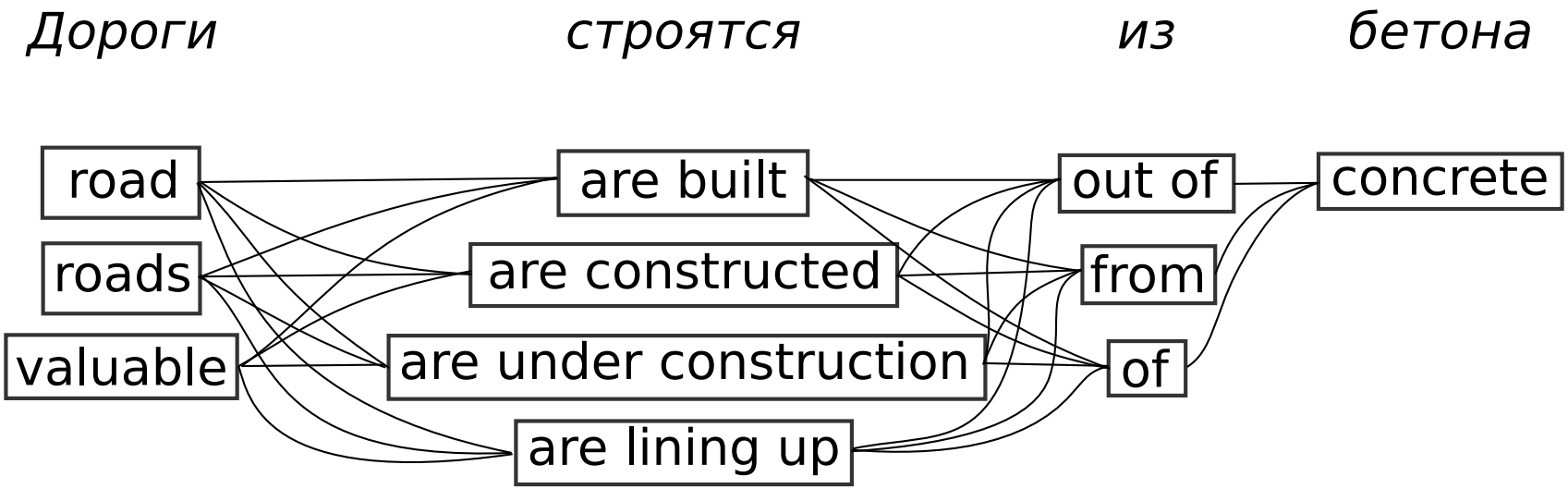
Английский со словарем

Пусть у нас есть словарь (т.е. перевод для каждого слова).

Дороги	*строятся*	*из*	*бетона*
road	are built	out of	concrete
roads	are constructed	from	
valuable	are under construction	of	
	are lining up		

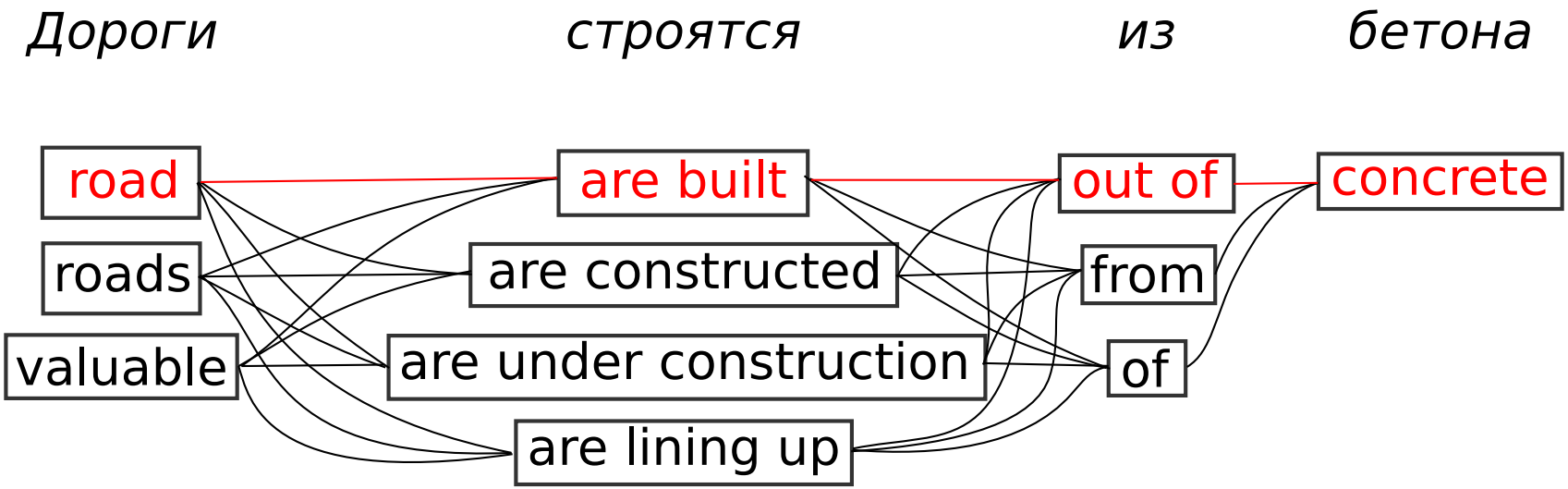
Английский со словарем

Пусть у нас есть словарь (т.е. перевод для каждого слова).



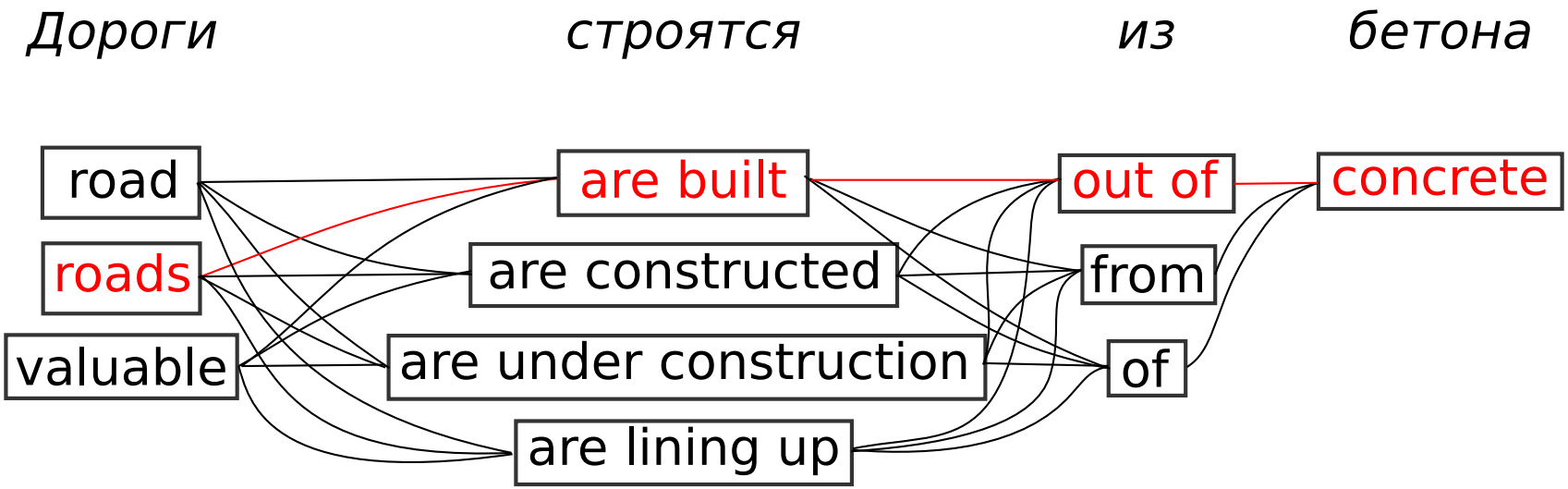
Английский со словарем

Пусть у нас есть словарь (т.е. перевод для каждого слова).



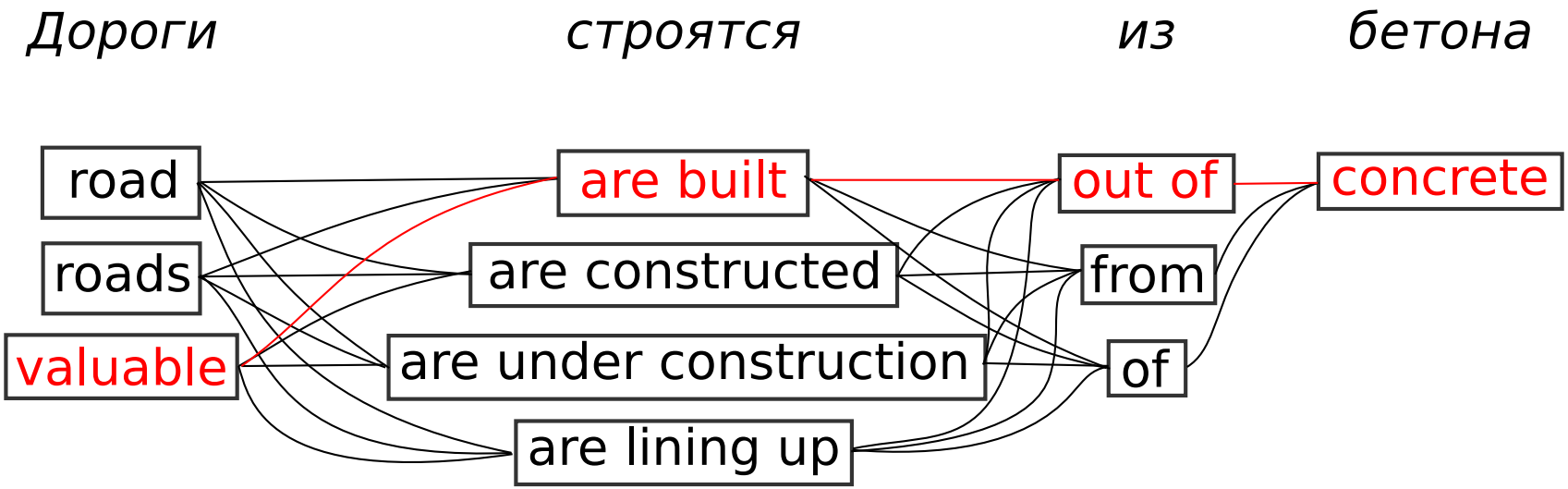
Английский со словарем

Пусть у нас есть словарь (т.е. перевод для каждого слова).



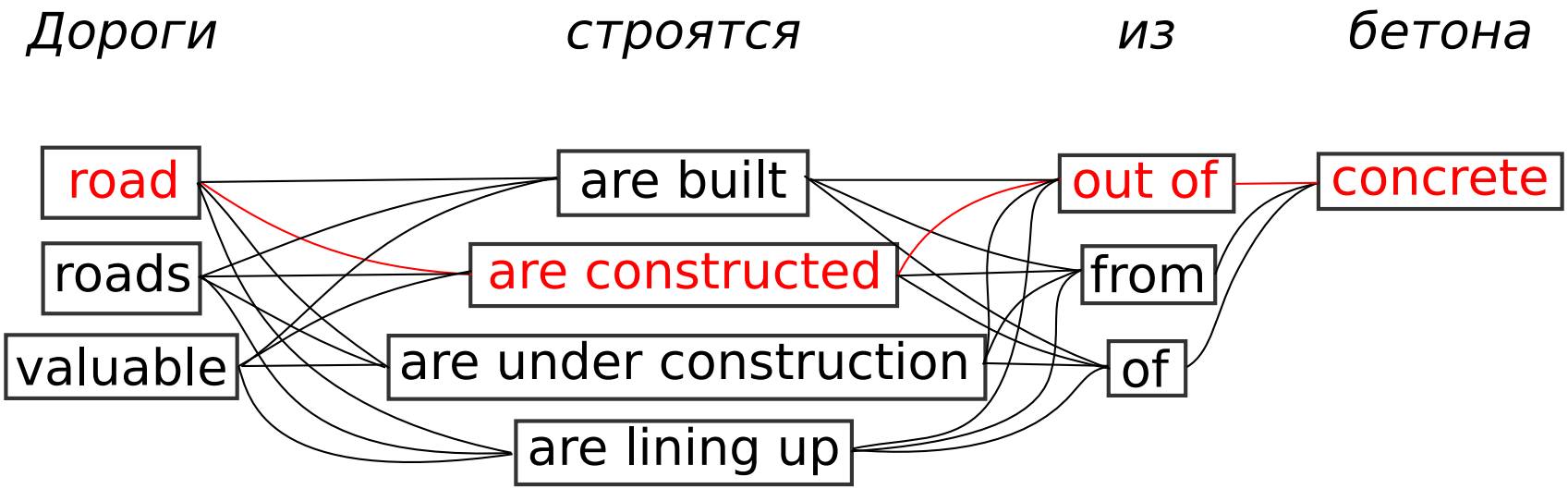
Английский со словарем

Пусть у нас есть словарь (т.е. перевод для каждого слова).



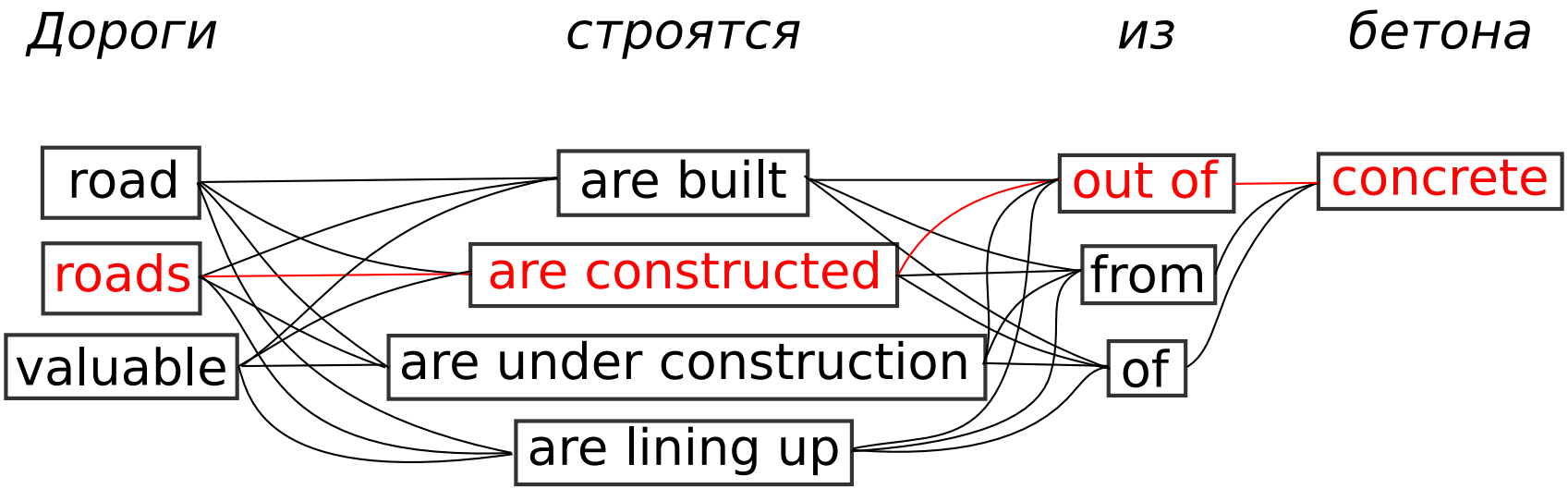
Английский со словарем

Пусть у нас есть словарь (т.е. перевод для каждого слова).



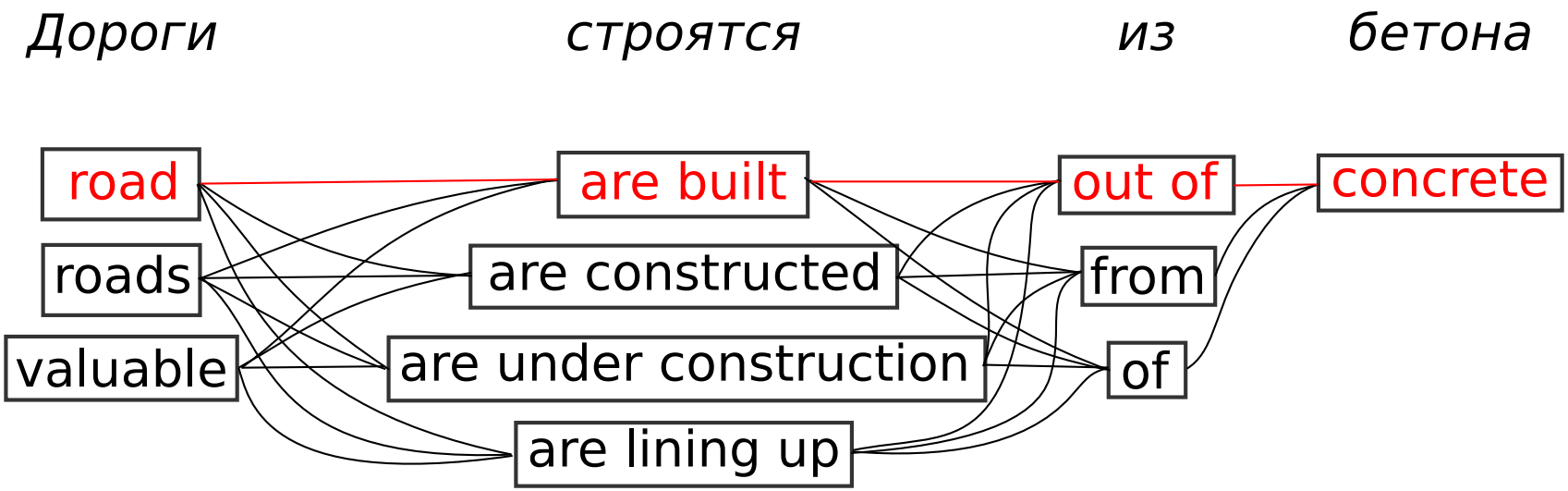
Английский со словарем

Пусть у нас есть словарь (т.е. перевод для каждого слова).



Английский со словарем

Пусть у нас есть словарь (т.е. перевод для каждого слова).

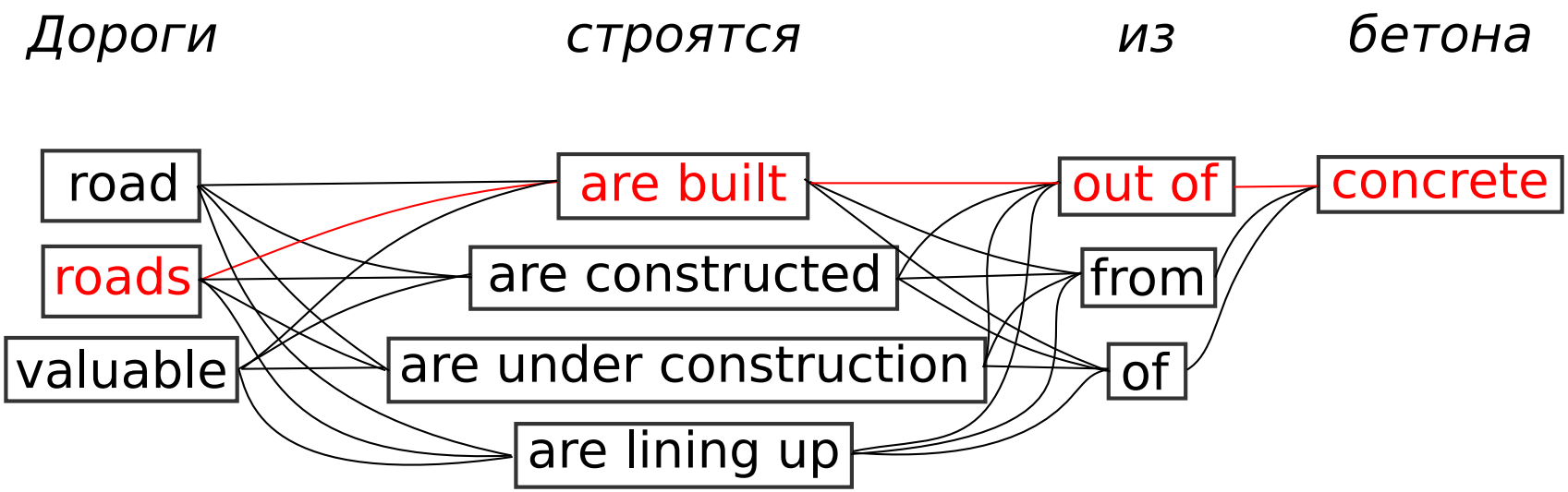


G(road are built out of concrete) = -40.7

Перебираем все пути и выбираем лучший.

Английский со словарем

Пусть у нас есть словарь (т.е. перевод для каждого слова).

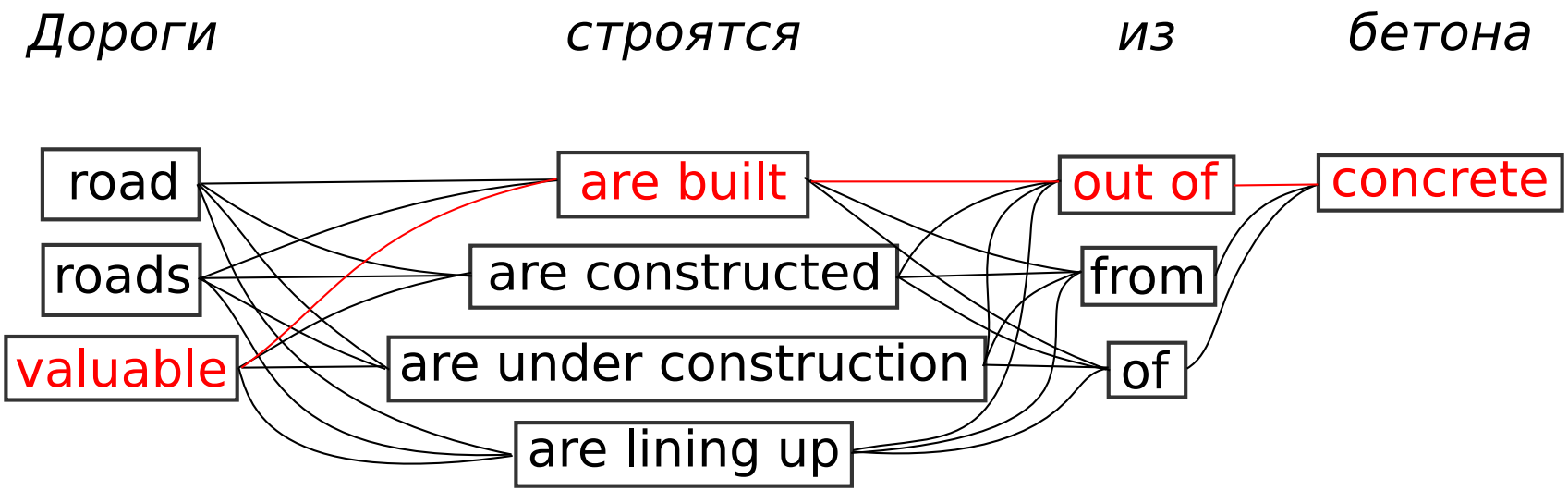


G(roads are built out of concrete) = -38.6

Перебираем все пути и выбираем лучший.

Английский со словарем

Пусть у нас есть словарь (т.е. перевод для каждого слова).

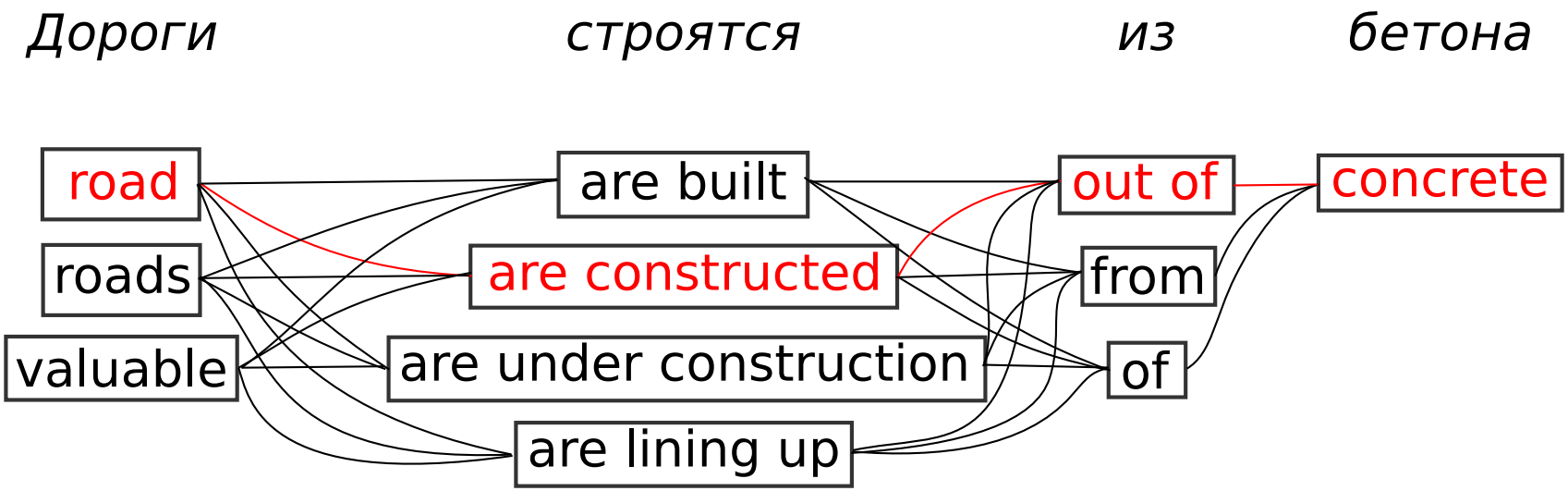


G(valuable are built out of concrete) = -45.7

Перебираем все пути и выбираем лучший.

Английский со словарем

Пусть у нас есть словарь (т.е. перевод для каждого слова).

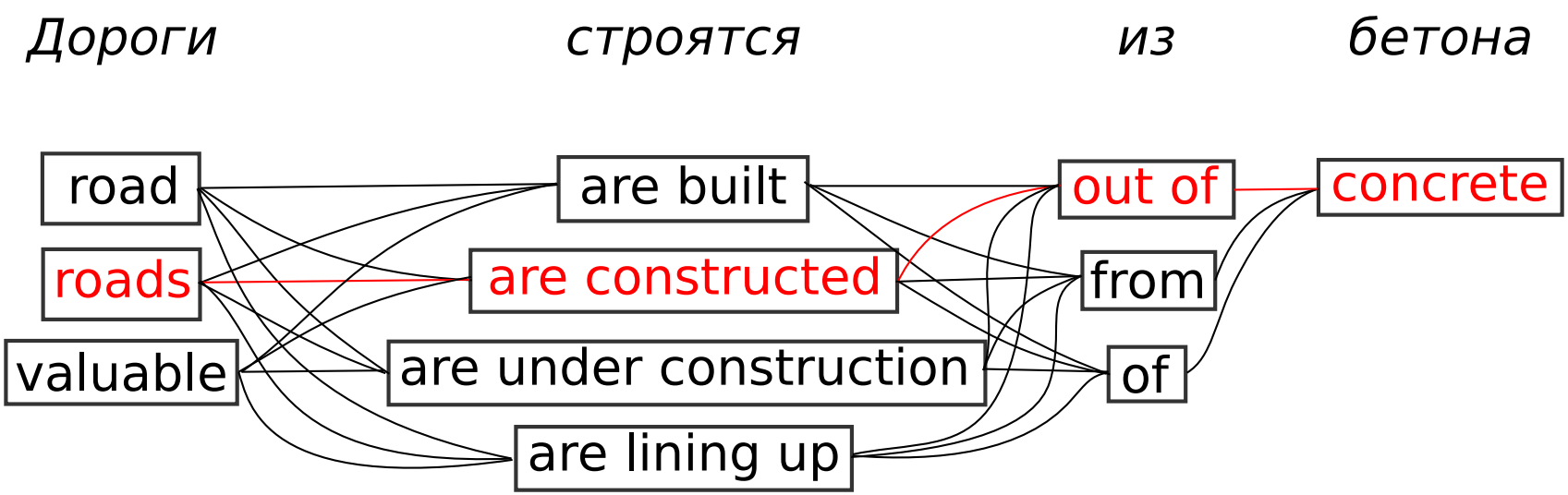


G(road are constructed out of concrete) = -41.1

Перебираем все пути и выбираем лучший.

Английский со словарем

Пусть у нас есть словарь (т.е. перевод для каждого слова).

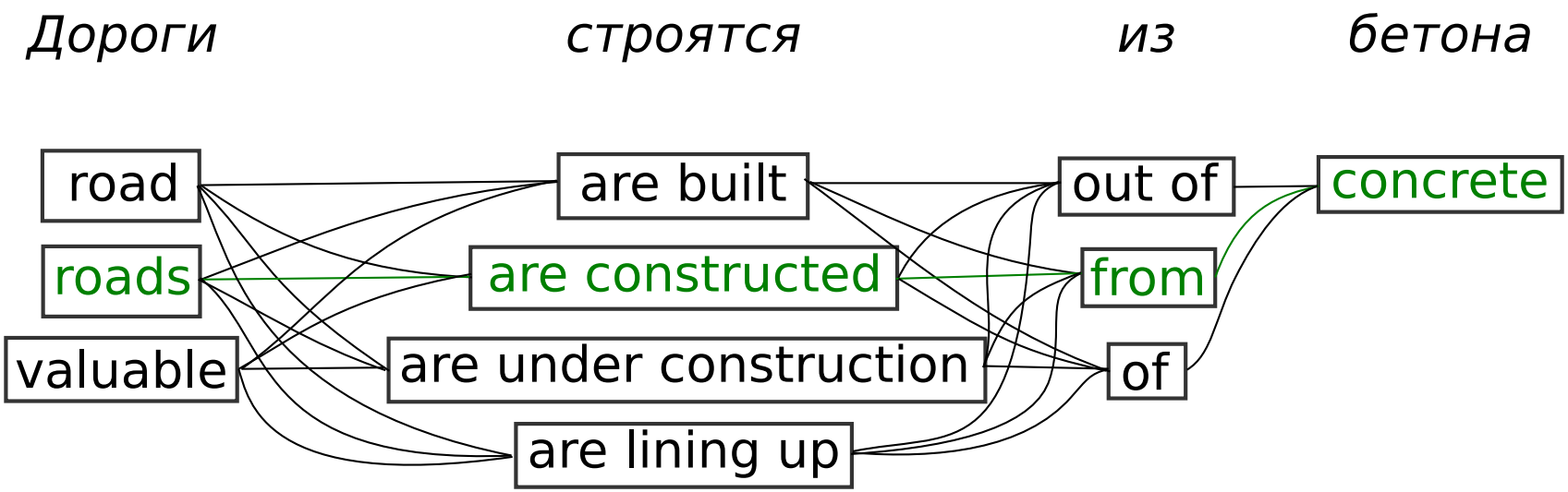


G(roads are constructed out of concrete) = -38.7

Перебираем все пути и выбираем лучший.

Английский со словарем

Пусть у нас есть словарь (т.е. перевод для каждого слова).



G(roads are constructed from concrete) = -32.6

Перебираем все пути и выбираем лучший.

Мера хорошести

Функция, которая принимает на вход предложение и выдает число *(меру хорошести)*.

- G(roads are built from concrete) = -34.9
- G(valuable are lining up among concrete) = -59.0

Выдаем перевод с наибольшей хорошестью.

Языковая модель

Чем вероятнее появление предложения в ~~языке~~ корпусе, тем больше число.

- LM(I'm fine, thanks.) = -26.1
- LM(Ash nazg durbatulûk, ash nazg gimbatul) = -31.3
- LM(tensorflow session thread-safe) = -126.9

Готово

У вас есть перевод.

(Осталось откуда-то взять словарь и функцию хорошести.)

Выравнивание

1. Выравнивание по документам

   `https://ru.wikipedia.org/wiki/Yandex`
   `https://en.wikipedia.org/wiki/Yandex`

Выравнивание

1. Выравнивание по документам

   `https://ru.wikipedia.org/wiki/Yandex`
   `https://en.wikipedia.org/wiki/Yandex`

2. Выравнивание по предложениям

Выравнивание

1. Выравнивание по документам

   `https://ru.wikipedia.org/wiki/Yandex`
   `https://en.wikipedia.org/wiki/Yandex`

2. Выравнивание по предложениям
3. Выравнивание по словам

Выравнивание по словам

Centauri and Arcturan languages.

ok-voon ororok sprok .	at-voon bichat dat .
ok-drubel ok-voon anok plok sprok .	at-drubel at-voon pippat rrat dat .
erok sprok izok hihok ghirok .	totat dat arrat vat hilat .
ok-voon anok drok brok jok .	at-voon krat pippat sat lat .
wiwok farok izok stok .	totat jjat quat cat .
lalok sprok izok jok stok .	wat dat krat quat cat .
lalok brok anok plok nok .	iat lat pippat rrat nnat .
wiwok nok izok kantok ok-yurp .	totat nnat quat oloat at-yurp .
alok mok nok yorok ghirok clok .	wat nnat gat mat bat hilat .
lalok nok crrrok hihok yorok zanzanok .	wat nnat arrat mat zanzanat .
lalok rarok nok izok hihok mok .	wat nnat forat arrat vat gat .

(Example by K. Knight, 1997)

Выравнивание по словам

Centauri and Arcturan languages.

ok-voon ororok sprok .	at-voon bichat dat .
ok-drubel ok-voon anok plok sprok .	at-drubel at-voon pippat rrat dat .
erok sprok izok hihok ghirok .	totat dat arrat vat hilat .
ok-voon anok drok brok jok .	at-voon krat pippat sat lat .
wiwok farok izok stok .	totat jjat quat cat .
lalok sprok izok jok stok .	wat dat krat quat cat .
lalok brok anok plok nok .	iat lat pippat rrat nnat .
wiwok nok izok kantok ok-yurp .	totat nnat quat oloat at-yurp .
alok mok nok yorok ghirok clok .	wat nnat gat mat bat hilat .
lalok nok crrrok hihok yorok zanzanok .	wat nnat arrat mat zanzanat .
lalok rarok nok izok hihok mok .	wat nnat forat arrat vat gat .

(Example by K. Knight, 1997)

Выравнивание по словам

Centauri and Arcturan languages.

ok-voon ororok sprok .	at-voon bichat dat .
ok-drubel ok-voon anok plok sprok .	at-drubel at-voon pippat rrat dat .
erok sprok izok hihok ghirok .	totat dat arrat vat hilat .
ok-voon anok drok brok jok .	at-voon krat pippat sat lat .
wiwok farok izok stok .	totat jjat quat cat .
lalok sprok izok jok stok .	wat dat krat quat cat .
lalok brok anok plok nok .	iat lat pippat rrat nnat .
wiwok nok izok kantok ok-yurp .	totat nnat quat oloat at-yurp .
alok mok nok yorok ghirok clok .	wat nnat gat mat bat hilat .
lalok nok crrrok hihok yorok zanzanok .	wat nnat arrat mat zanzanat .
lalok rarok nok izok hihok mok .	wat nnat forat arrat vat gat .

(Example by K. Knight, 1997)

Выравнивание по словам

Centauri and Arcturan languages.

ok-voon ororok sprok .	at-voon bichat dat .
ok-drubel ok-voon anok plok sprok .	at-drubel at-voon pippat rrat dat .
erok sprok izok hihok ghirok .	totat dat arrat vat hilat .
ok-voon anok drok brok jok .	at-voon krat pippat sat lat .
wiwok farok izok stok .	totat jjat quat cat .
lalok sprok izok jok stok .	wat dat krat quat cat .
lalok brok anok plok nok .	iat lat pippat rrat nnat .
wiwok nok izok kantok ok-yurp .	totat nnat quat oloat at-yurp .
alok mok nok yorok ghirok clok .	wat nnat gat mat bat hilat .
lalok nok crrrok hihok yorok zanzanok .	wat nnat arrat mat zanzanat .
lalok rarok nok izok hihok mok .	wat nnat forat arrat vat gat .

(Example by K. Knight, 1997)

Выравнивание по словам

Centauri and Arcturan languages.

ok-voon ororok sprok .	at-voon bichat dat .
ok-drubel ok-voon anok plok sprok .	at-drubel at-voon pippat rrat dat .
erok sprok izok hihok ghirok .	totat dat arrat vat hilat .
ok-voon anok drok brok jok .	at-voon krat pippat sat lat .
wiwok farok izok stok .	totat jjat quat cat .
lalok sprok izok jok stok .	wat dat krat quat cat .
lalok brok anok plok nok .	iat lat pippat rrat nnat .
wiwok nok izok kantok ok-yurp .	totat nnat quat oloat at-yurp .
alok mok nok yorok ghirok clok .	wat nnat gat mat bat hilat .
lalok nok crrrok hihok yorok zanzanok .	wat nnat arrat mat zanzanat .
lalok rarok nok izok hihok mok .	wat nnat forat arrat vat gat .

(Example by K. Knight, 1997)

Выравнивание по словам

Centauri and Arcturan languages.

ok-voon ororok sprok .	at-voon bichat dat .
ok-drubel ok-voon anok plok sprok .	at-drubel at-voon pippat rrat dat .
erok sprok izok hihok ghirok .	totat dat arrat vat hilat .
ok-voon anok drok brok jok .	at-voon krat pippat sat lat .
wiwok farok izok stok .	totat jjat quat cat .
lalok sprok izok jok stok .	wat dat krat quat cat .
lalok brok anok plok nok .	iat lat pippat rrat nnat .
wiwok nok izok kantok ok-yurp .	totat nnat quat oloat at-yurp .
alok mok nok yorok ghirok clok .	wat nnat gat mat bat hilat .
lalok nok crrrok hihok yorok zanzanok .	wat nnat arrat mat zanzanat .
lalok rarok nok izok hihok mok .	wat nnat forat arrat vat gat .

(Example by K. Knight, 1997)

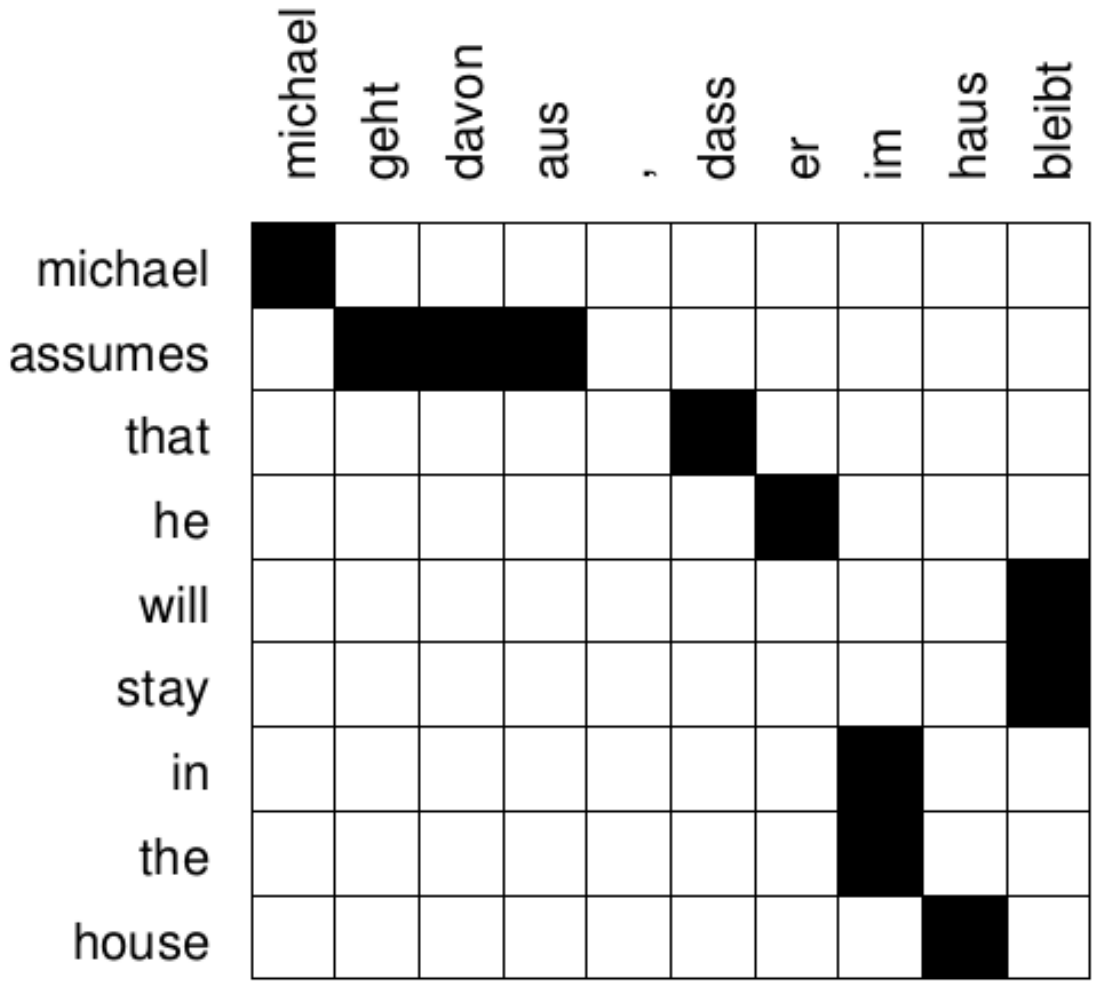
Выравнивание по словам

Centauri and Arcturan languages.

ok-voon ororok sprok .	at-voon bichat dat .
ok-drubel ok-voon anok plok sprok .	at-drubel at-voon pippat rrat dat .
erok sprok izok hihok ghirok .	totat dat arrat vat hilat .
ok-voon anok drok brok jok .	at-voon krat pippat sat lat .
wiwok farok izok stok .	totat jjat quat cat .
lalok sprok izok jok stok .	wat dat krat quat cat .
lalok brok anok plok nok .	iat lat pippat rrat nnat .
wiwok nok izok kantok ok-yurp .	totat nnat quat oloat at-yurp .
alok mok nok yorok ghirok clok .	wat nnat gat mat bat hilat .
lalok nok crrrok hihok yorok zanzanok .	wat nnat arrat mat zanzanat .
lalok rarok nok izok hihok mok .	wat nnat forat arrat vat gat .

(Example by K. Knight, 1997)

Выравнивание по словам



(From slides by P. Koehn)

Данные

- Классификация – 150 примеров.

Данные

- Классификация – 150 примеров.
- Penn treebank (English) – 50K предложений.

Данные

- Классификация – 150 примеров.
- Penn treebank (English) – 50K предложений.
- ImageNet – 15M картинок.

Данные

- Классификация – 150 примеров.
- Penn treebank (English) – 50K предложений.
- ImageNet – 15M картинок.
- FR-EN: **30M** параллельных предложений, **1B** EN monolingual предложений в открытом доступе (у поисковиков в **100 раз** больше).

Данные

Крупнейшая *supervised* задача, для которой данные для обучения *лежат просто так:*

- Интернет
  - `https://ru.wikipedia.org/wiki/Yandex`
  - `https://en.wikipedia.org/wiki/Yandex`
- Субтитры
- Parliament proceedings (European, Canadian, etc.), законы
- Википедия
- Книги, песни, инструкции, и т.д.

План

История

Задача

Качество

Треугольник

Структура курса

Метрика качества перевода

Надо:

$$f(r_{1..R}, t_{1..T}) \to [0, 1],$$

где r – reference, t – translation.

- Показывать, насколько хорошо перевод сохраняет информацию
- Показывать, насколько перевод "гладкий"
- Не зависеть от языка
- Учитывать порядок слов
- ...

Отсутствие единственного эталона

这个机场的安全工作由以色列方面负责。

- Israeli officials are responsible for airport security.
- Israel is in charge of the security at this airport.
- The security work for this airport is the responsibility of the Israel government.
- Israeli side was in charge of the security of this airport.
- Israel is responsible for the airport's security.
- Israel is responsible for safety work at this airport.
- Israel presides over the security of the airport.
- Israel took charge of the airport security.
- The safety of this airport is taken charge of by Israel.
- This airport's security is the responsibility of the Israeli security officials.

(From slides by P. Koehn)

BLEU

$$
\begin{aligned}
BLEU(s,t) &= bp \cdot \sqrt[4]{p_1 \cdot p_2 \cdot p_3 \cdot p_4} \\
p_n &= \text{modified ngram precision} \\
bp &= \min(1, \exp(1 - R/T))
\end{aligned}
$$

BLEU

$r =$ the cat is on the mat
$t =$ the the kitty is on the mat

BLEU

r = the cat is on the mat
t = the the kitty is on the mat

Translation:
1-gr: the, the, kitty, is, on, the, mat
2-gr: the the, the kitty, kitty is, is on, on the, the mat
3-gr: the the kitty, kitty is on, is on the, on the mat
4-gr: the the kitty is, the kitty is on, kitty is on the, is on the mat

BLEU

r = the cat is on the mat
t = the the kitty is on the mat

Translation:
1-gr: the, the, kitty, is, on, the, mat
2-gr: the the, the kitty, kitty is, is on, on the, the mat
3-gr: the the kitty, kitty is on, is on the, on the mat
4-gr: the the kitty is, the kitty is on, kitty is on the, is on the mat

BLEU

r = the cat is on the mat
t = the the kitty is on the mat

Translation:
1-gr: the, the, kitty, is, on, the, mat
2-gr: the the, the kitty, kitty is, is on, on the, the mat
3-gr: the the kitty, kitty is on, is on the, on the mat
4-gr: the the kitty is, the kitty is on, kitty is on the, is on the mat

$p_1 = 5/7, p_2 = 3/6, p_3 = 2/5; p_4 = 1/4; bp = 1$

$BLEU = 0.43$

Качество перевода

HTER	**assessment**	**application examples**
0%		Seamless bridging of language divide
	publishable	Automatic publication of official announcements
10%		
	editable	Increased productivity of human translators
20%		Access to official publications
		Multi-lingual communication (chat, social networks)
30%	gistable	Information gathering
		Trend spotting
40%	triagable	Identifying relevant documents
50%		

(Tables by P. Koehn)

Качество перевода

HTER	**assessment**	**language pairs and domains**
0%		
	publishable	French-English restricted domain
10%		French-English tecnhical document localization
	editable	French-English news stories
20%		
		French-German news stories
30%	gistable	English-Czech open domain
40%	triagable	
50%		

(Tables by P. Koehn)

План

История

Задача

Качество

Треугольник

Структура курса

Треугольник

Лексический трансфер

- `John read the letter`
- `John ga tegami o yon-da`

Лексический трансфер

- John read the letter
- John ga tegami o yon-da
- John

Лексический трансфер

- John read the letter
- John ga tegami o yon-da
- John yon-da

Лексический трансфер

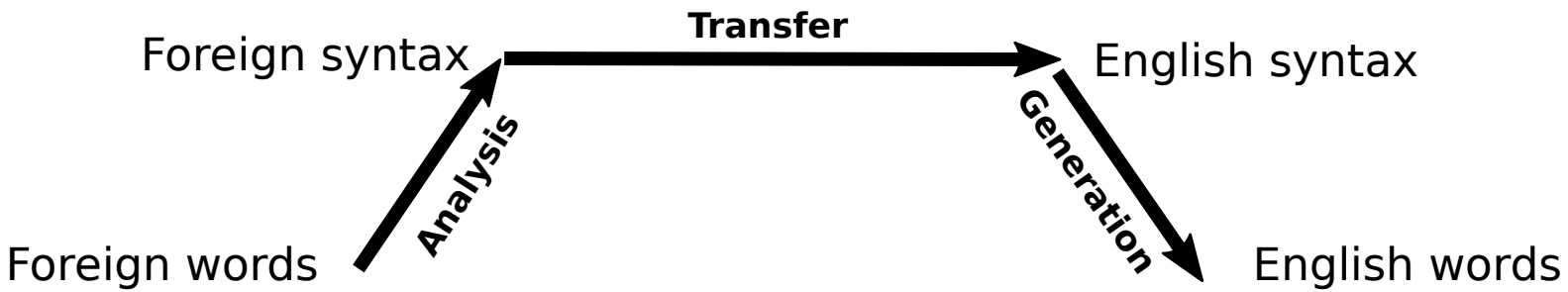
- John read the letter
- John ga tegami o yon-da
- John yon-da tegami

Лексический трансфер

- John read the letter
- John ga tegami o yon-da
- John yon-da tegami

Треугольник

Треугольник



Синтаксический трансфер

- John read the letter
- John ga tegami o yon-da
-

Синтаксический трансфер

- John read the letter
- John ga tegami o yon-da
- *John read the letter*

Синтаксический трансфер

- John read the letter
- John ga tegami o yon-da
- *John the letter read*

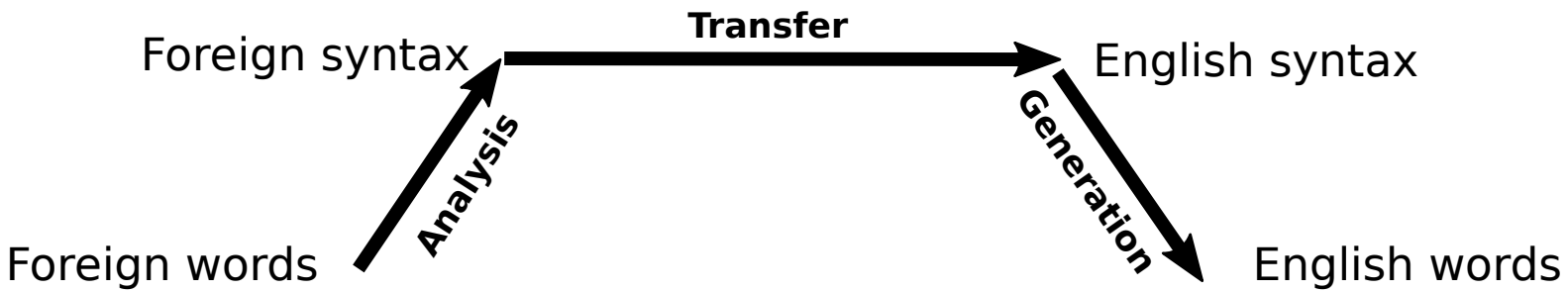
Синтаксический трансфер

- John read the letter
- John ga tegami o yon-da
- *John* ga *the letter* o *read*

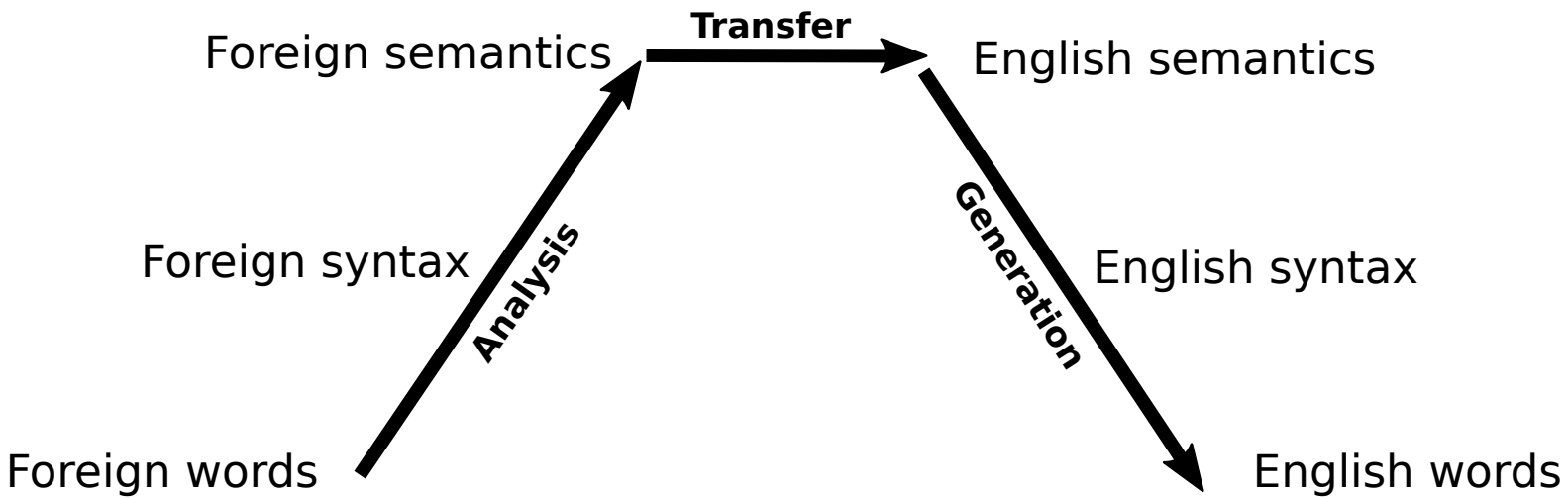
Синтаксический трансфер

- John read the letter
- John ga tegami o yon-da
- John ga tegami o yon-da

Треугольник



Треугольник



Семантический трансфер

- John read the letter
- John ga tegami o yon-da
-

```
Выполняется действие
    Действие:           read
    Кто выполняет:      John
    Над чем выполняет:  the letter
    Когда выполняет:    (неизвестно)
```

Семантический трансфер

- John read the letter
- John ga tegami o yon-da

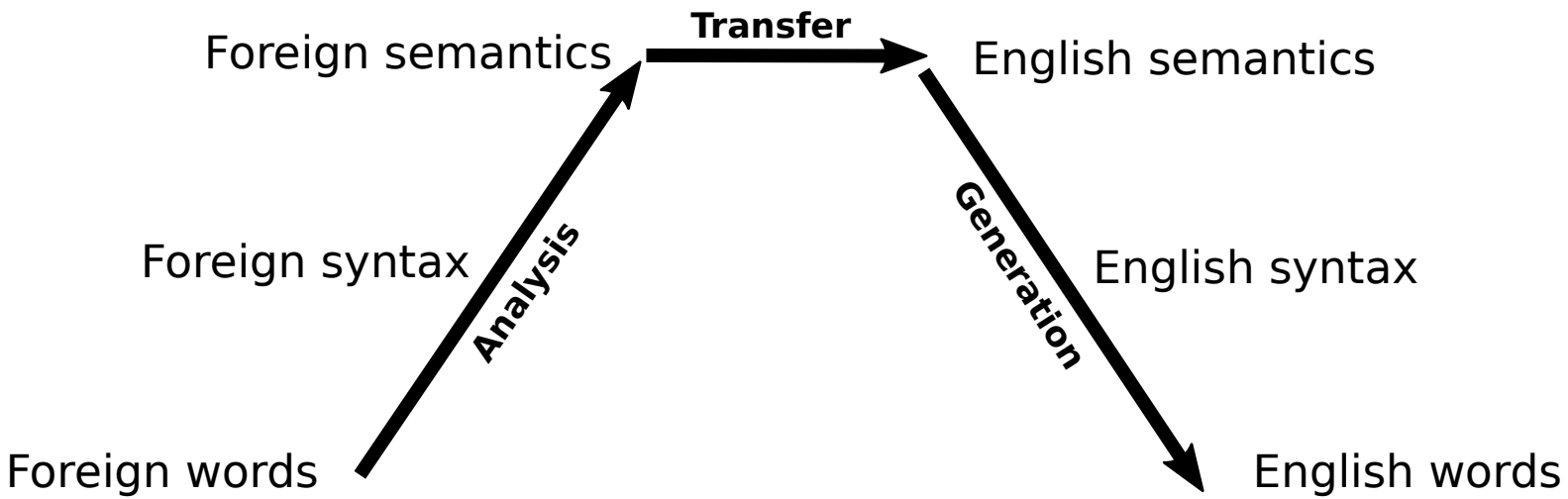
```
Выполняется действие
    Действие:              yon-da
    Кто выполняет:         John
    Над чем выполняет:     tegami
    Когда выполняет:       (неизвестно)
```

Семантический трансфер

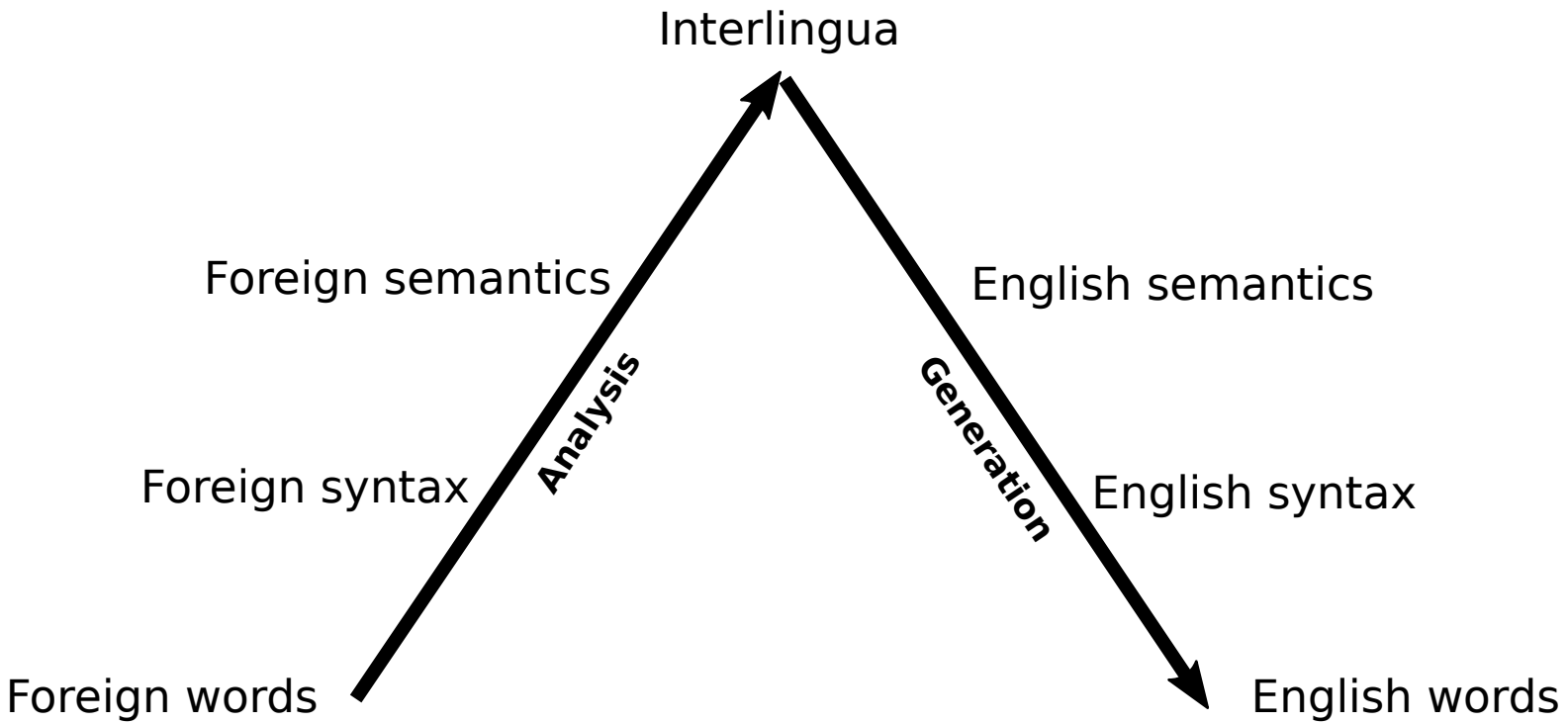
- John read the letter
- John ga tegami o yon-da
- John ga tegami o yon-da

```
Выполняется действие
    Действие:              yon-da
    Кто выполняет:         John
    Над чем выполняет:     tegami
    Когда выполняет:       (неизвестно)
```

Треугольник



Треугольник



Интерлингва

- John read the letter
- John ga tegami o yon-da
-

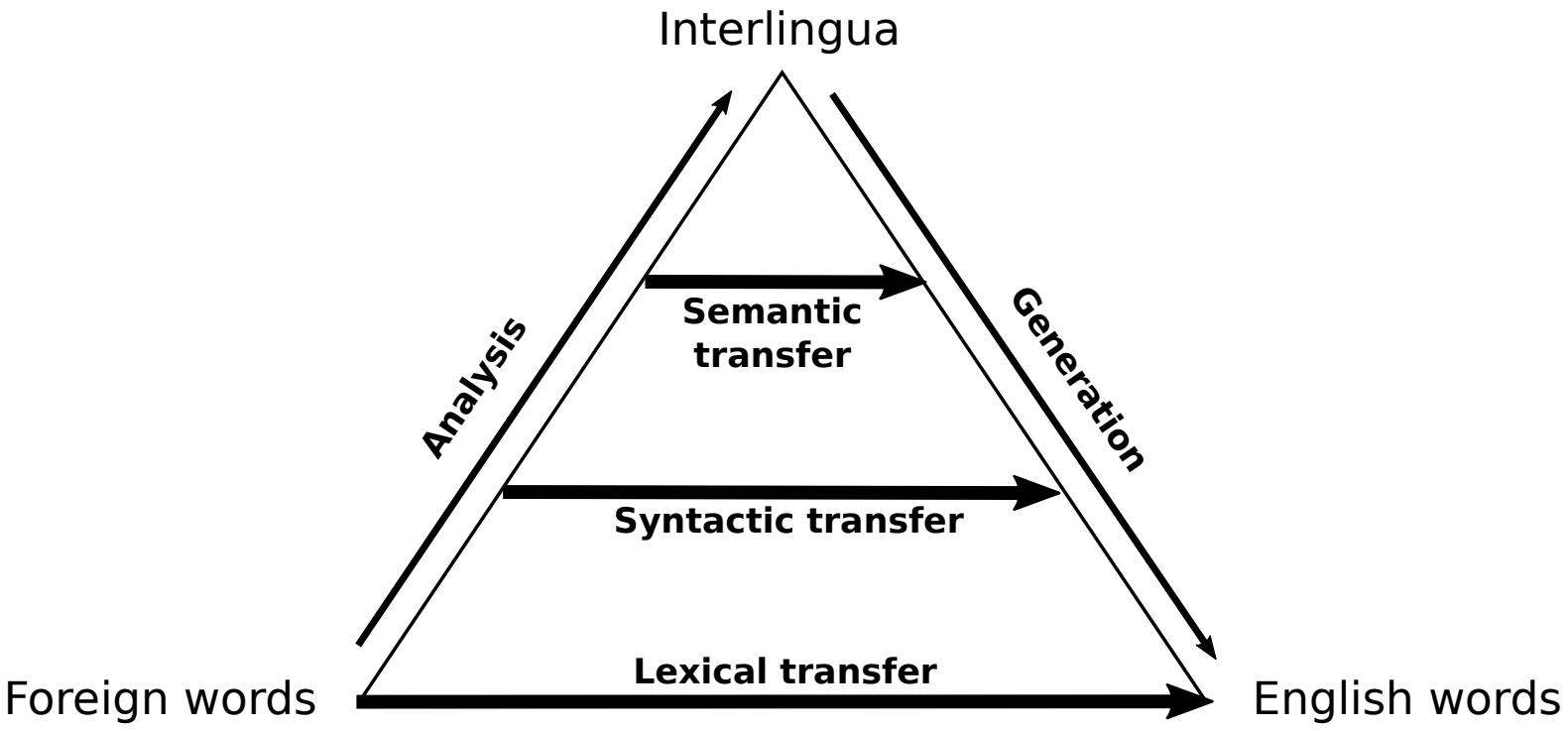
Интерлингва

- John read the letter
- John ga tegami o yon-da
- [0.539, 0.150, -0.012, ...]

Интерлингва

- John read the letter
- John ga tegami o yon-da
- John ga tegami o yon-da

Треугольник



План

История

Задача

Качество

Треугольник

Структура курса

Структура курса

https://wiki.school.yandex.ru/shad/groups/2015/Semester4/MachineTranslate

- 6 statistical machine translation lectures (3 PBMT + 3 NMT).
- 4 rule-based machine translation lectures.