Как работи машинният превод?

Проф. д-р Светла Коева
Институт за български език
„Проф. Любомир Андрейчин“
Българска академия на науките

Машинен превод?

- През 2020 година:
- 44 зетабайта информация в интернет на различни езици;
- по 6 компютърни устройства на жител на планетата.

Машинен превод с правила

- Базира се на граматични и семантични правила за анализ и синтез на словосъчетания и изречения.
- Необходими са изчерпателни речници и правила, включващи подробна морфологична, синтактична и семантична информация за двойката езици, за които се осъществява преводът.
- Подобни ресурси обикновено се изготвят от квалифицирани специалисти и изработването им е сложна и продължителна задача.

Статистически машинен превод

- Статистическият машинен превод се базира на вероятността дадена последователност от символи да е превод на друга последователност от символи.
- При статистическия машинен превод се използват статистически модели, които се създават на базата на тренировъчни корпуси.
- Тренировъчните корпуси са паралелни преводни текстове и текстове от езика, на който се превежда.

Трудности

- Езиковият знак е арбитрарен (случаен).

zaldi	*igel*	*txoria*	*behi*	*sagua*

http://grammar.about.com/od/ab/g/Arbitrariness.htm

Трудности

- Езиковият знак е арбитрарен (случаен).

zaldi	*igel*	*txoria*	*behi*	*sagua*
horse	*frog*	*bird*	*cow*	*mouse*

Трудности

- Граматичността не винаги означава приемливост.

Colorless green ideas sleep furiously.

****Furiously sleep ideas green colorless.***

Chomsky. N. Syntactic Structures. The Hague/Paris: Mouton. 1957.

Трудности

Бе сгладне и честлинните комбурси тарляха се и сврецваха във плите; съвсем окласни бяха тук щурпите и отма равапсатваха прасурси.

Луис Карол. Алиса в Страната на чудесата и в Огледалния свят. София, 1996.

Трудности

- Многозначност

***Старите** приятели и познати на Ана*

*си спомниха нейната младост **в Пловдив**.*

Трудности

- Междуезикова асиметрия

  - The German chancellor Angela Merkel will make an announcement on Thursday.
  - Angela Merkel wird am Donnerstag eine Ankündigung machen.
  - Германският канцлер Ангела Меркел ще направи изявление в четвъртък.

Как работи машинният превод?

изход

вход

$$e_{best} = \arg\max_e P(e|f)$$
$$= \arg\max_e P(f|e)P(e)$$

Алгоритъм

Преводен модел

Езиков модел

Статистически машинен превод

- Статистическият МТ се обучава от данни:
  - преводни текстове;
  - текстове от езика, на който се превежда.

- Колкото повече данни, толкова по-добре.
- Колкото повече подходящи данни, толкова по-добре.

Какво може да се научи от данните?

- Кое изречение в кое изречение се превежда: съотнасяне по изречения.
- Коя дума с коя дума се превежда: съотнасяне по думи и преводна вероятност.
- Дали полученият превод е лексикално, граматически и стилистично коректен: езиков модел.

Съотнасяне по изречения

	Английски	Български
ID01	INFORMATION FOR NATIONAL PARLIAMENTS	ИНФОРМАЦИЯ ЗА НАЦИОНАЛНИТЕ ПАРЛАМЕНТИ
ID02	Article 1	Член 1
ID03	Commission consultation documents (green and white papers and communications) shall be forwarded directly by the Commission to national Parliaments upon publication.	Консултативните документи на Комисията (зелени книги, бели книги и съобщения) се изпращат директно от Комисията на националните парламенти на държавите членки при публикуването им.
ID04	The Commission shall also forward the annual legislative programme as well as any other instrument of legislative planning or policy to national Parliaments, at the same time as to the European Parliament and the Council.	Комисията изпраща на националните парламенти и годишната законодателна програма, както и всеки друг инструмент за законодателно планиране или за политическа стратегия, като едновременно с това ги изпраща и на Европейския парламент и на Съвета.

Съотнасяне по думи

I had dropped — бях изпуснал

the lantern — фенера

when I seized — когато се хванах

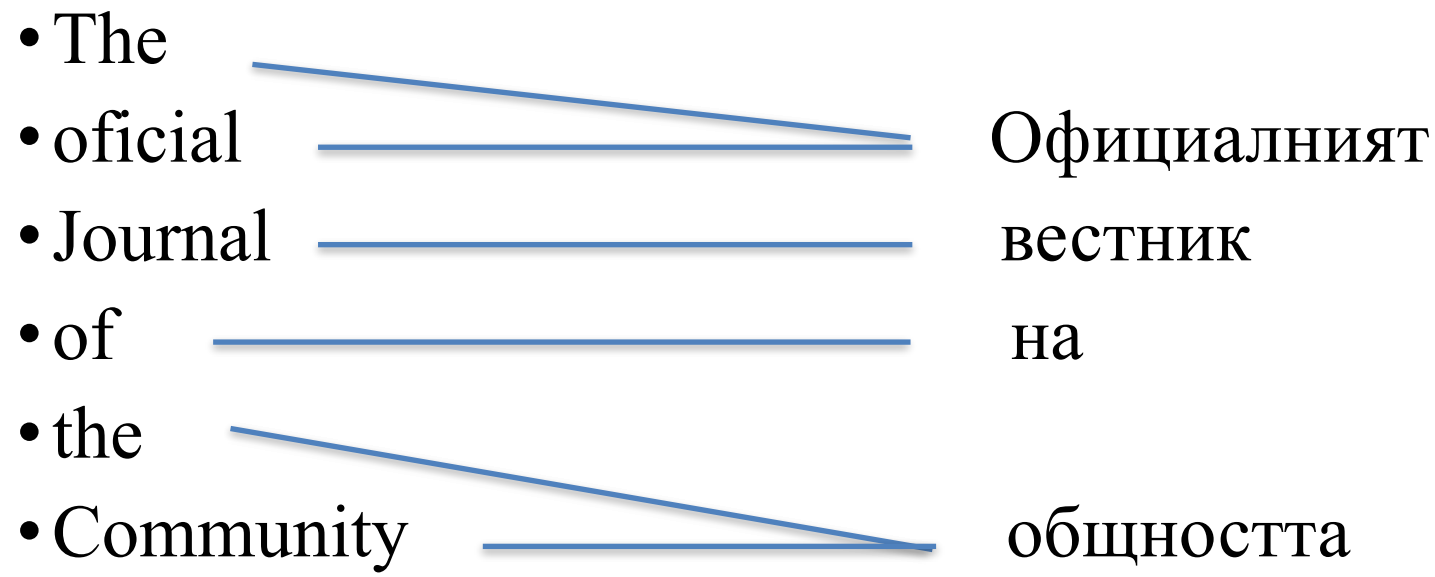
the bars — [за] решетката

but it still — но той все още

burned upon the floor — гореше на пода

http://www.bglibrary.net/en-bg-title.htm

Съотнасяне по думи



Съотнасяне по думи

酸辣汤	**Люто кисела супа**
鲜蘑牛肉汤	**Супа от телешко и печурки**
鲜蘑鸡肉汤	**Супа от пиле с печурки**
鲜蘑鸡	**Пиле с печурки**

Статистически машинен превод

I love this boy
Аз обичам това момче
I love this dog
Аз обичам това куче
They love this dog
Те обичат това куче
I talk to this dog
Аз говоря на кучето

Съотнесени думи и изречения

Статистически машинен превод

I love this boy
Аз обичам това момче
I love this dog
Аз обичам това куче
They love this dog
Те обичат това куче
I talk to this dog
Аз говоря на кучето

I	Аз	3
love	обичам	2
love	обичат	1
this	това	3
this		1
boy	момче	1
dog	куче	2
dog	кучето	1
They	Те	1
talk	говоря	1
talk	говорят	1

Статистически машинен превод

I love this boy
Аз обичам това момче
I love this dog
Аз обичам това куче
They love this dog
Те обичат това куче
I talk to this dog
Аз говоря на кучето

I	Аз	3/3
love	обичам	2/3
love	обичат	1/3
this	това	3/4
this		1/4
boy	момче	1/1
dog	куче	2/3
dog	кучето	1/3
They	Те	1/1
talk	говоря	1/1
talk	говорят	1/1

Статистически машинен превод

I love this boy
Аз обичам това момче
I love this dog
Аз обичам това куче
They love this dog
Те обичат това куче
I talk to this dog
Аз говоря на кучето

They love this boy.

They	Те	1/1
love	обичам	2/3
love	обичат	1/3
this	това	3/4
this		1/4
boy	момче	1/1

Езиков модел

Статистически машинен превод

I love this boy
Аз обичам това момче
I love this dog
Аз обичам това куче
They love this dog
Те обичат това куче
I talk to this dog
Аз говоря на кучето

They love this boy.		
They	Те	1/1
love	обичат	1/3
this	това	3/4
boy	момче	1/1

Езиков модел

Какво е правилно в езика, на който се превежда.

Превод, базиран на фрази

- По-добър модел:
  - Не се превежда дума по дума;
  - Съотнасянето по фрази позволява по-коректно да се представят редица езикови черти, например съгласуване.

- this boy : това момче
- to this dog : на кучето
- I talk : Аз говоря

Статистически машинен превод

I love this boy
Аз обичам това момче
I love this dog
Аз обичам това куче
They love this dog
Те обичат това куче
I talk to this dog
Аз говоря на кучето

I love	Аз обичам	2
They love	Те обичат	1
this boy	това момче	1
this dog	това куче	2
I talk	Аз говоря	1
to	на	1
this dog	кучето	1

Статистически машинен превод

I love this boy
Аз обичам това момче
I love this dog
Аз обичам това куче
They love this dog
Те обичат това куче
I talk to this dog
Аз говоря на кучето

I love	Аз обичам	2
They love	Те обичат	1
this boy	това момче	1
this dog	това куче	2
I talk	Аз говоря	1
to	на	1
this dog	кучето	1

They love this dog.

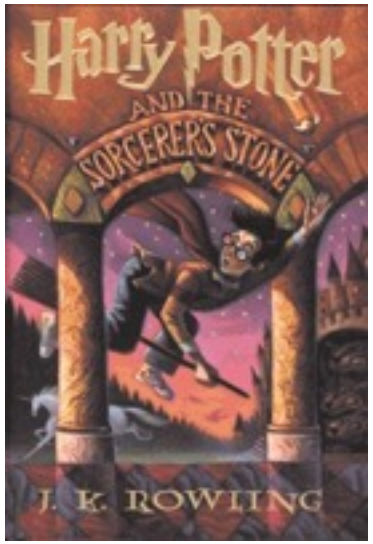
Те обичат това куче.

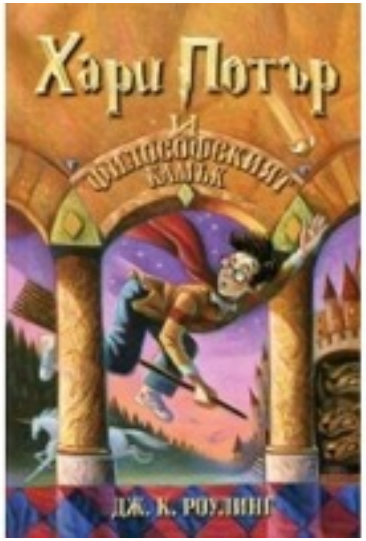
Те обичат кучето.

Машинният превод и данните

- Статистическият машинен превод зависи от данните.
- Статистическият машинен превод се учи как да превежда от данните.
- Данните са:
  - преводни паралелни текстове;
  - едноезикови текстове (за езика, на който се превежда) ;
  - речници, терминологични бази, онтологии, имена на лица, организации, географски имена.
- Също като при хората статистическият машинен превод е толкова по-добър, колкото повече се е учил.

Машинният превод и данните

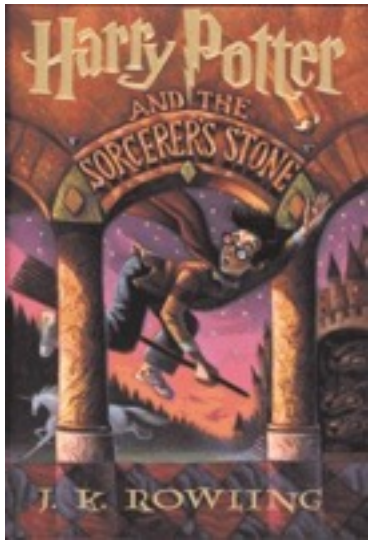


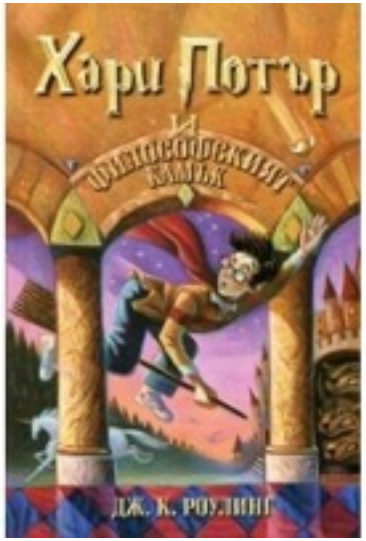


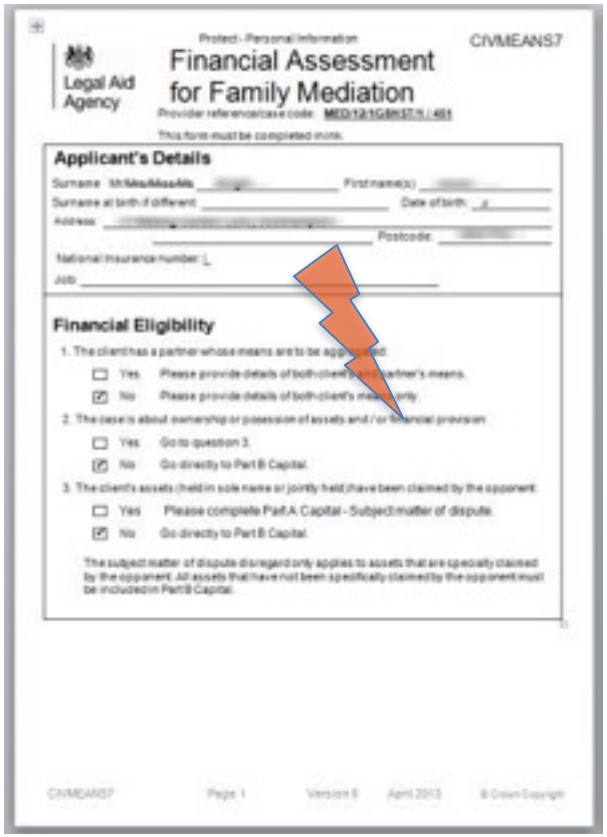

"Good," said Wizard Wink. "We'll start with Numbers."
"Oh no!" groaned Max and Mick.
The wizard boys didn't like school much. They wanted to play.
Wizard Wink asked them the first question.
"If I had three crystal balls in one hand . . ." he said, "and four in the other . . . what would I have?"
"BIG HANDS!" shouted Max and Mick.
"Very funny," said Wizard Wink.
"Oooh! I know. I know!" said Zoe. "Seven!"
"Right!" said Wizard Wink.
"Poo!" said Melissa. "That was easy-peasy."
That was quick!
But it wasn't easy!

MOSES CORE

Машинният превод и данните







MOSES CORE

Платформата CEF.AT

- Механизмът за свързване на Европа осигурява услуги за гражданите на Европейския съюз.
- *Платформата за автоматичен превод CEF.AT* се нуждае от подходящи данни, за да работи по-добре.
- Свободният софтуер Мозес е научна разработка, финансирана от Европейската комисия, и се използва от MT@EC – част от *CEF.AT* .

Координация на езиковите ресурси в Европа

Да се подкрепят нашите езици означава да се подкрепя обединена Европа и обратно.