Jezikovne tehnologije

Strojno prevajanje

Jernej Vičič

Vsebina

1. Uvod
2. Strojno prevajanje
3. Korpusni pristop
4. Statistično strojno prevajanje
5. Primeri

1 Uvod

- brez uvoda

2 Strojno prevajanje

• Machine translation (MT) is the application of computers to the task of translating texts from one natural language to another

• (FAMT) Fully Automatic Machine Translation translation of natural languages with no user intervention

(http://www.eamt.org/)

2 Strojno prevajanje, zgodovina

- začetek,
- prva leta,
- petdeseta leta prejšnjega stoletja,
- osemdeseta in začetki devetdesetih,
- zdaj.

2 Strojno prevajanje, zgodovina

- 1700 in prej: Leibniz in Descartes,
- "translating machines", trak z besedami,
- pravi začetki digitalnega MT,
- petdeseta leta,
  - Georgetown-IBM experiment,
- ALPAC report (1966),
- sodobni sistemi za strojno prevajanje.

2 Strojno prevajanje, zgodovina

The Good News According to Mark:

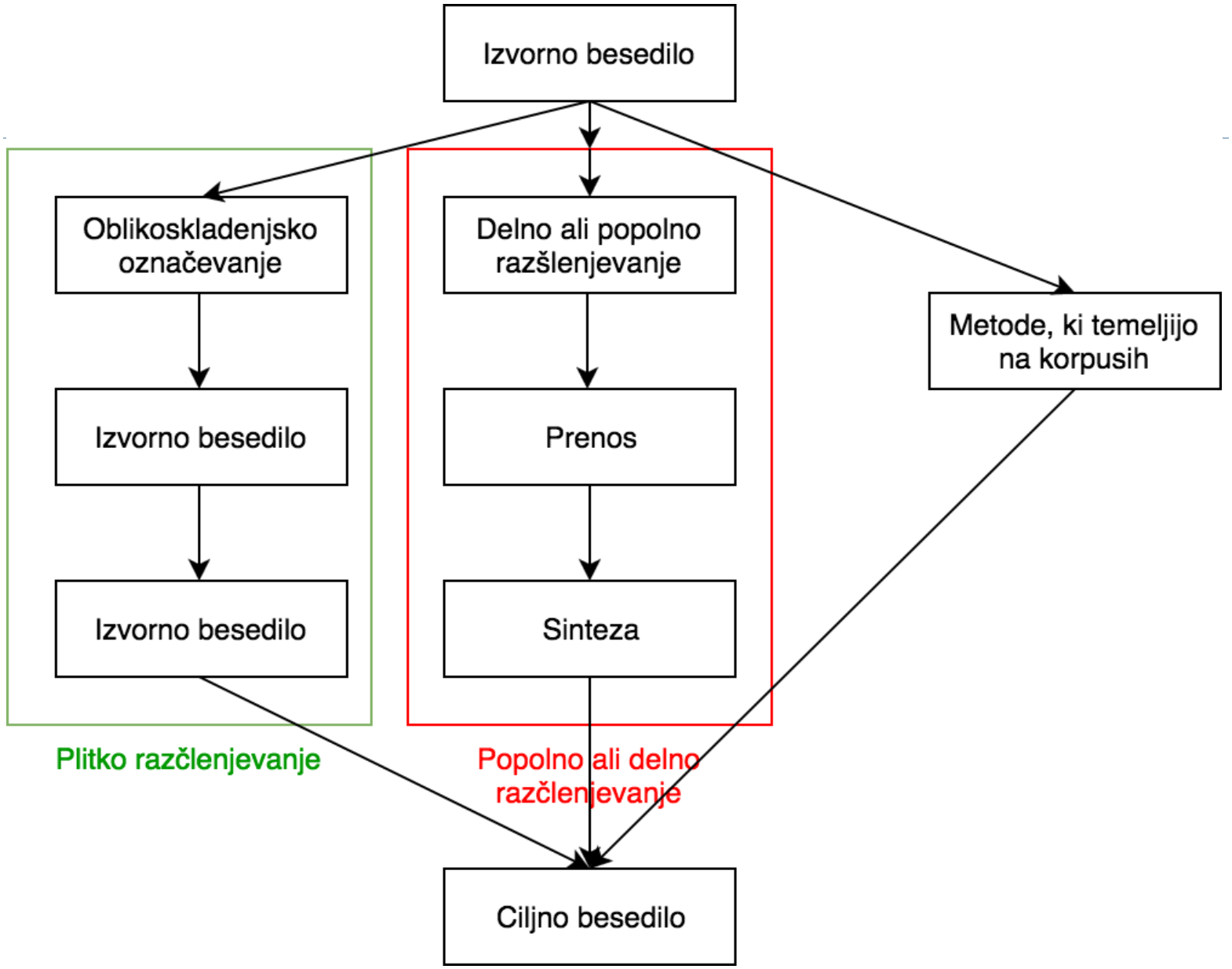
“The spirit indeed is willing, but the flesh is weak.”

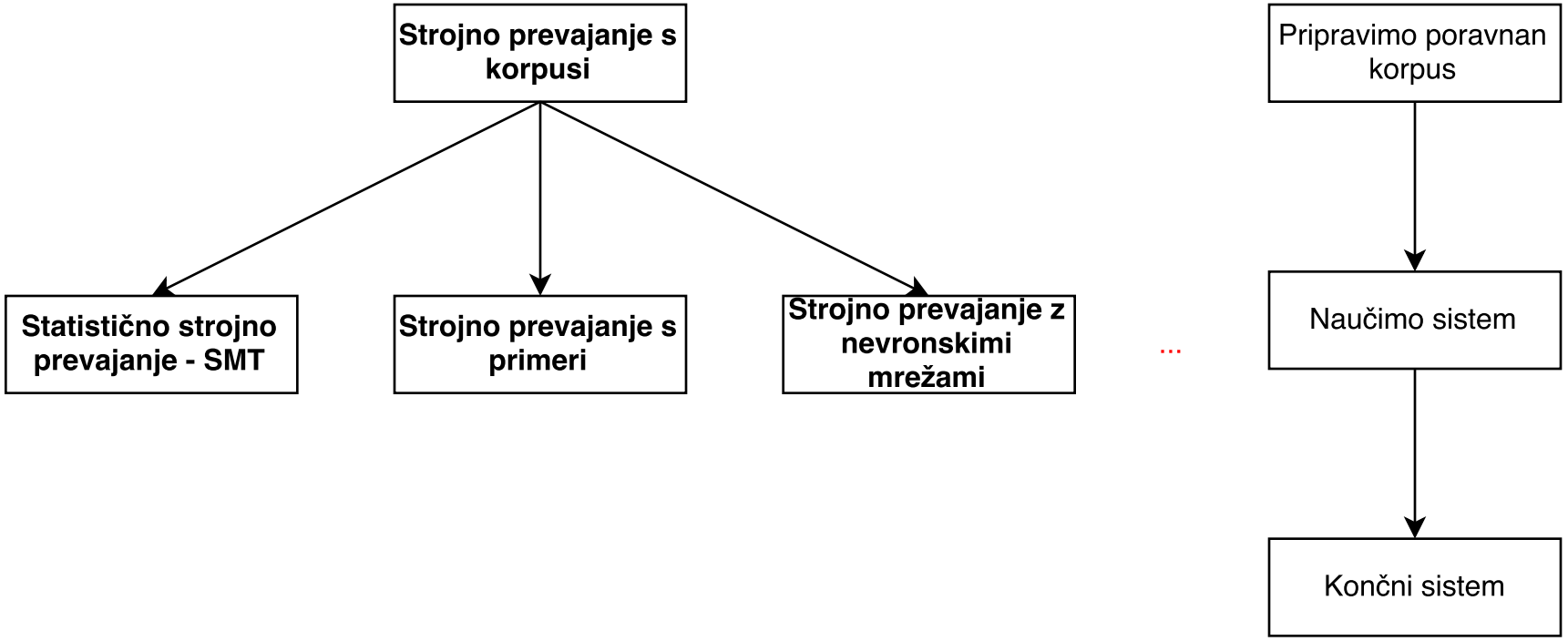
prevod:

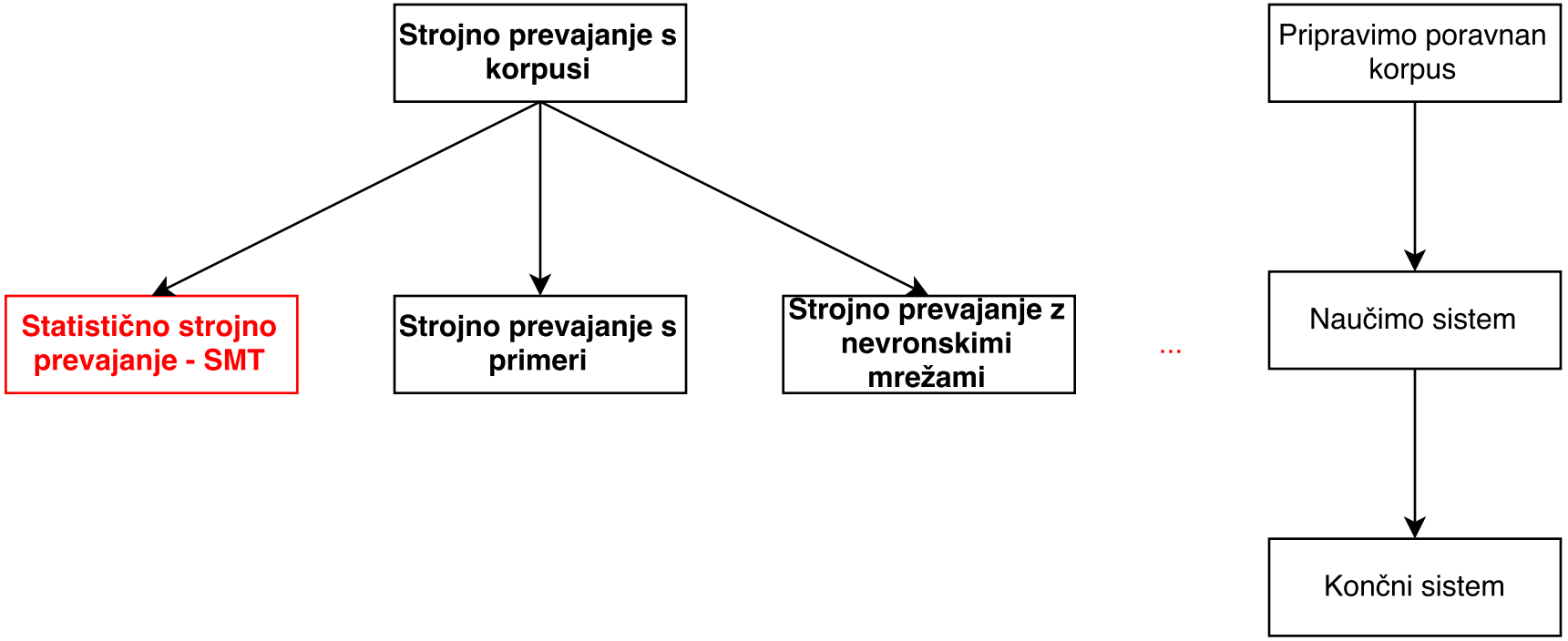
“The vodka is good, but the flesh is rotten.”

2 Strojno prevajanje, razdelitev

- strojno prevajanje – SP,
- SP na osnovi pravil (Rule-based MT),
- SP na osnovi korpusov (Corpus-based MT),
- statistično SP (Statistical MT),
- SP na osnovi primerov (Example-based MT),
- hibridno SP (Hybrid MT),
- SP na osnovi nevronskih mrež (Neural machine translation NMT).


Izvorno besedilo
Oblikoskladenjsko označevanje
Delno ali popolno razšlenjevanje
Metode, ki temeljijo na korpusih
Izvorno besedilo
Prenos
Izvorno besedilo
Sinteza
Plitko razčlenjevanje
Popolno ali delno razčlenjevanje
Ciljno besedilo


Strojno prevajanje s korpusi
Statistično strojno prevajanje - SMT
Strojno prevajanje s primeri
Strojno prevajanje z nevronskimi mrežami
...
Pripravimo poravnan korpus
Naučimo sistem
Končni sistem


Strojno prevajanje s korpusi
Statistično strojno prevajanje - SMT
Strojno prevajanje s primeri
Strojno prevajanje z nevronskimi mrežami
...
Pripravimo poravnan korpus
Naučimo sistem
Končni sistem

2 Strojno prevajanje, korpusni pristop

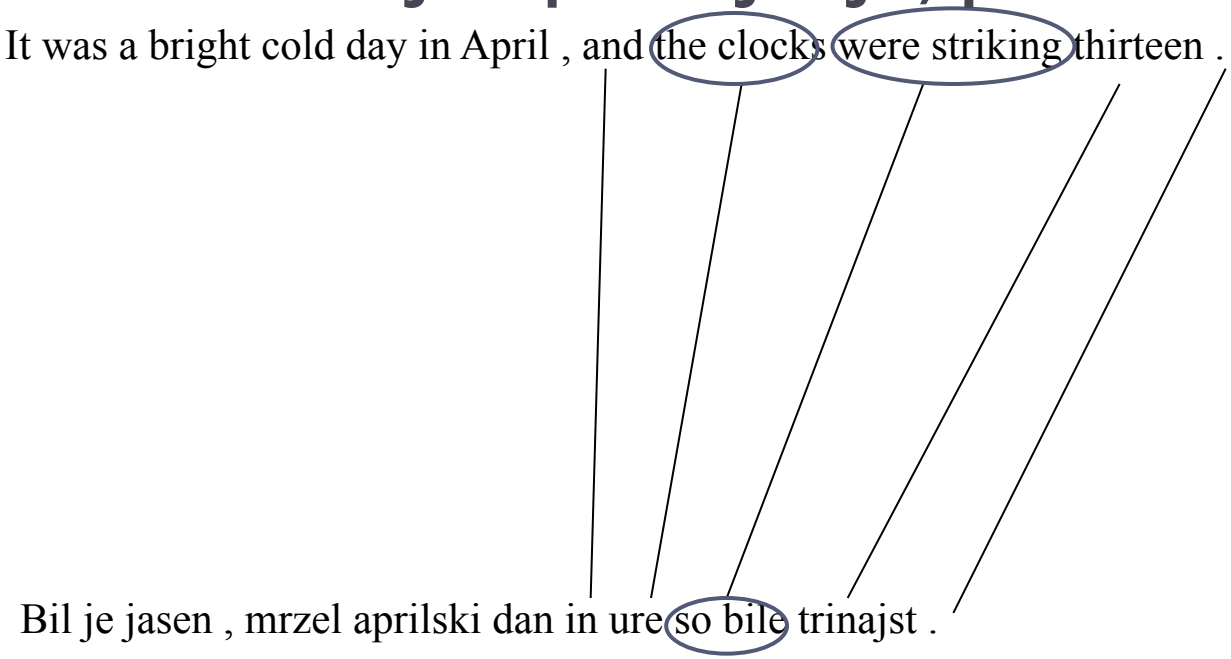
- korpus,
- dvojezični korpus,
- poravnani korpus.

2 Strojno prevajanje, korpusni pristop

1. It was a bright cold day in April , and the clocks were striking thirteen .
2. Winston Smith , his chin nuzzled into his breast in an effort to escape the vile wind , slipped quickly through the glass doors of Victory Mansions , though not quickly enough to prevent a swirl of gritty dust from entering along with him .
3. The hallway smelt of boiled cabbage and old rag mats .
4. At one end of it a coloured poster , too large for indoor display , had been tacked to the wall .
5. It depicted simply an enormous face , more than a metre wide : the face of a man of about forty-five , with a heavy black moustache and ruggedly handsome features .

1. Bil je jasen , mrzel aprilski dan in ure so bile trinajst .
2. Winston Smith je imel brado zakopano v prsi , da bi ušel strupenemu vetru , ko je stopil skozi steklena vrata bloka Zmaga , vendar ne dovolj hitro , da ne bi vrtinec peščenega prahu vstopil skupaj z njim .
3. Veža je smrdela po kuhanem zelju in starih , cunjastih predpražnikih .
4. Na eni strani je bil na steno pribit barven , za notranjo opremo prevelik plakat .
5. Prikazoval je preprosto ogromen , več kot meter velik obraz : obraz moškega pri petinštiridesetih , s košatimi črnimi brki in z ostro začrtanimi , čednimi potezami .

2 Strojno prevajanje, primeri poravnav



2 Strojno prevajanje, primeri poravnav

Kupil bom lepo obleko .

I will buy a nice dress .

Jaz vozim rdeč avtomobil .

I drive a red car .

Kaj si mi kupil ?

What did you buy me ?

2 Strojno prevajanje, korpusni pristop

- potrebni veliki poravnani korpusi,
- težko nadzorujemo,
- majhen vložek (ob pripravljenih korpusih),
- trenutno najbolj zanimivo področje

2 Strojno prevajanje, SMT

- temelji na verjetnosti,
- temelji na velikih količinah primerov,
- matematično „lepi“ modeli,
- (zaenkrat) se ne zaveda okolice,
- rezultate težko preverjamo (zakaj),
- napake težko odpravljamo.

2 Strojno prevajanje, SMT

- Predstavniki:
- Google translate (NI VEČ!!!),
- Microsoft BING translator,
- IBM (Brown in sodelavci),
- Moses (ogrodje za postavitev lastnega sistema),
- Menola (moja malenkost – na podlagi št. 3).

2 Strojno prevajanje, SMT

- potrebujemo velik dvojezični korpus:
- poravnane povedi izvornega ter ciljnega jezika;
- izdelamo prevajalni model.
- Potrebujemo velik enojezični korpus:
  - izdelamo jezikovni model ciljnega jezika.

2 Strojno prevajanje, SMT

- učna faza,
- prevajalna faza.

2 Strojno prevajanje, SMT

- vzporedna besedila (prevodi) – korpus,
- po povedih,
- nekaj milijonov,
- Google pravi vsaj 100 milijonov besed,
- povedi dolge 10 – 20 besed,
- enojezični korpus ciljnega jezika,
- nekaj 10 milijonov besed (raje več) – Google milijardo besed.

Verjetnost

- poskus: n med seboj enakih izidov,
- dogodek A: m ugodnih izidov,
  - –Verjetnost dogodka A:

$$P(A) = \frac{m}{n} = \frac{število\_ugodnih\_izidov}{število\_vseh\_izidov}$$

- primer:
  - –v tej učilnici je n poslušalcev,
  - –m poslušalcev ima očala,
  - –izračunaj verjetnost, da ima izbrani poslušalec očala!

Verjetnost

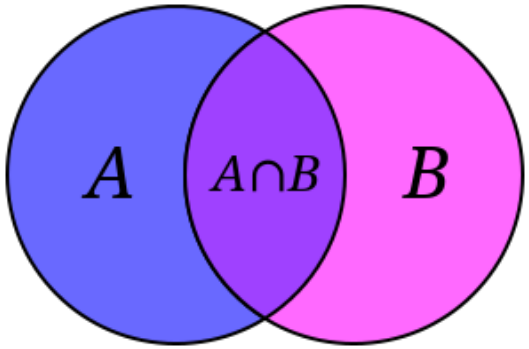
- n=? (20),
- m=? (5),

$$P(A) = \frac{m}{n} = \frac{5}{20} = 0{,}25 = 25\%$$

Pogojna verjetnost

- verjetnost, da se zgodi dogodek A,
  - –pod pogojem, da se je zgodil dogodek B.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(AB)}{P(B)}$$



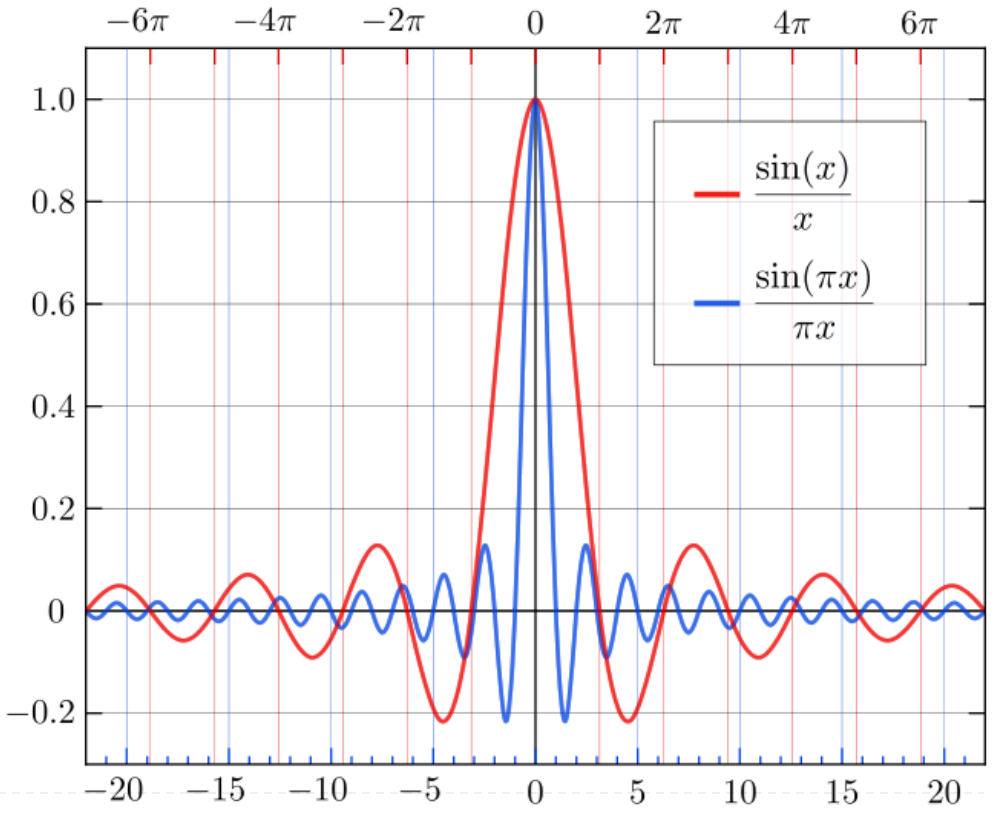
- Pazi P(B)!
  - •inženirji izberemo dovolj majhno število

argmax

- argmax - **arguments of the maxima**

$argmax(\frac{\sin \tau x}{\tau x})=0$

$argmax(\frac{\sin x}{x})=0$



SMT – formule

e – ciljna poved
f – izvorna poved
$\hat{e}$ – prevedena poved

iščemo: $\hat{e} = \arg\max_{e} P(e \mid f)$

Gospod Bayes pa je rekel:

$$\hat{e} = \arg\max_{e} P(e \mid f) =$$

$$\arg\max_{e} P(f \mid e) P(e)$$

SMT

$$\hat{e} = \arg\max_{e} P(e \mid f) =$$

$$\arg\max_{e} P(f \mid e) P(e)$$

$P(f \mid e)$ - verjetnost za izvorno poved, pri pogoju, da ciljna obstaja

$P(e)$ - verjetnost za ciljno poved

Verjetnost

$P(f \mid e)$ -verjetnost za izvorno poved, pri pogoju, da ciljna obstaja,

-to je prevajalni model (translation model - TM),

$P(e)$ -verjetnost za ciljno poved,

-to je jezikovni model (language model - LM).

Prevajalni model

- dodeli višjo vrednost parom, ki se večkrat pojavljajo.

p(trinajst|thirteen)=0,02

p(in|and)=0,03

p(avto|car)=0,004

p(avto|vehicle)=0,002

p(in|banana)=0,0000001

Model ciljnega jezika

- ponavadi n-gram modeli:
  - n-gram: n-terica,
  - ponavadi trigram,
  - dodeli višjo vrednost povedim, ki so dobre v ciljnem jeziku.

p(v|sem šel)=0,001

p(srečen|bil sem)=0,0001

p(srečen|banana hruška)=0,00000001

Unigram modeli

- najbolj enostaven primer: $p(x_1 \ldots x_n) = \prod_{i=1}^{n} p(x_i)$

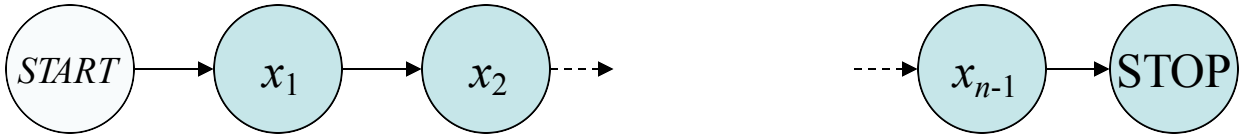
Narišemo:

Problem z unigrami: P(je je je je) >>>> P(Danes je lepo vreme)!

Bigram modeli

Verjetnost glede na prejšnjo besedo:

$$p(x_1 \dots x_n) = \prod_{i=1}^{n} p(x_i | x_{i-1})$$



n-gram modeli

Please close the door

Please close the first window on the left

198015222 the first
194623024 the same
168504105 the following
158562063 the world
…
14112454 the door

23135851162 the *

197302 close the window
191125 close the door
152500 close the gap
116451 close the thread
87298 close the deal

3785230 close the *

3380 please close the door
1601 please close the window
1164 please close the new
1159 please close the gate
…
0 please close the first

13951 please close the *

n-gram modeli

Točna dekompozija: verjetnostna porazdelitev:

$$p(x_1 \dots x_n) = p(x_1) \prod_{i=1}^{n} p(x_i | x_1 \dots x_{i-1})$$

(verjetnost glede na celotno zgodovino)

k-gram modeli (k>1): verjetnost na k-1 prejšnjih besed

$$p(x_1 \dots x_n) = \prod_{i=1}^{n} q\left(x_i | x_{i-(k-1)} \dots x_{i-1}\right)$$

Učna faza: določamo distribucijo: $q(x_i | x_{i-(k-1)} \dots x_{i-1})$

parametri modela

Parametri modela:

Maximum likelihood estimate: relativna frekvenca

$$q_{ML}(w) = \frac{c(w)}{c()}, \quad q_{ML}(w|v) = \frac{c(w,v)}{c(v)}, \quad q_{ML}(w|u,v) = \frac{c(w,u,v)}{c(u,v)}, \quad \ldots$$

c(v) – count – preštejemo število pojavitev v

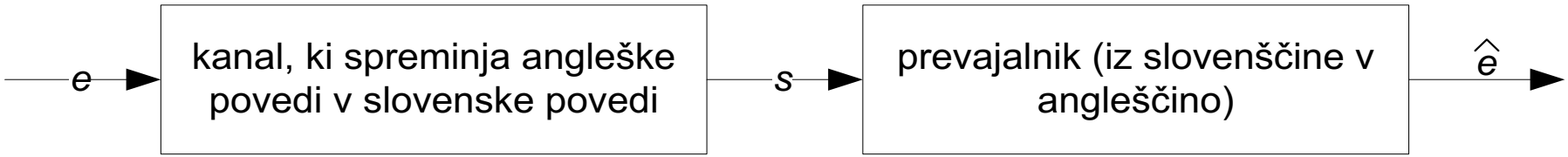
$$q(\text{avto}|\text{rdeč}) = \frac{14112454}{2313581162}$$

$$= 0.0006$$

SMT - Šumni kanal

- vzemimo primer prevajanja slovenskih povedi v angleške;
- za podano slovensko poved s si zamislimo, da je bila najprej zgrajena iz ekvivalentne angleške povedi e;
- angleška poved je potovala po šumnem kanalu s posebno lastnostjo: angleške povedi pretvori v slovenske;
- predpostavka: ta kanal lahko preizkusno določimo in matematično opišemo.

Šumni kanal



- e - namišljena angleška poved (izvorna),
- s - vhodna slovenska poved (to prevajamo v angleško),
- ê - angleški prevod vhodne povedi s.

Modeliranje

- Prevajalni modeli:
  - “adequacy”,
  - dodelijo boljše rezultate pravilnim (in celotnim) prevodom.
- Jezikovni modeli:
  - “fluency”,
  - dodelijo boljše rezultate besedilom v ciljnem jeziku.

Moses

- odprtokodna zbirka orodij,
- že pripravljeni modeli,
- lahko naredimo svoj prevajalni sistem.

5 Problemi s slovenščino

- miza
- mize
- mizi
- mizo
- mizi
- mizo

table

5 Kako ujamemo “zlikovca”?

- najbolj pogoste napake sistemov SMT:
  - prevod prek pivotnega jezika,
    - tuje besede,
  - ujemanje bližnjih besed v:
    - spolu, sklonu, številu,
    - samostalniška fraza.

5 Kako ujamemo “zlikovca”?

Today I bought a nice red car and I will go for a ride later this afternoon.

Danes sem kupil lepo rdeči avto in bom šel za vožnjo kasneje popoldne.

Danes	danes	Rsn
sem	biti	Gp-spe-n
kupil	kupiti	Ggdd-em
lepo	lepo	Rsn
rdeči	rdeč	Ppnmetd
avto	avto	Sometn
in	in	Vp
bom	biti	Gp-ppe-n
šel	iti	Ggvd-em
za	za	Dt
vožnjo	vožnja	Sozet
kasneje	kasno	Rsr
popoldne	popoldne	Rsn