A gépi fordítás alapjai

avagy

Gondotatok és metareflexiók a számítógépes nyelvészet témakörében, különös tekintettel a gépi fordítás szempontjaira

Oravecz Csaba

Számítógépes nyelvészet kurzus
ELTE Elméleti Nyelvészet

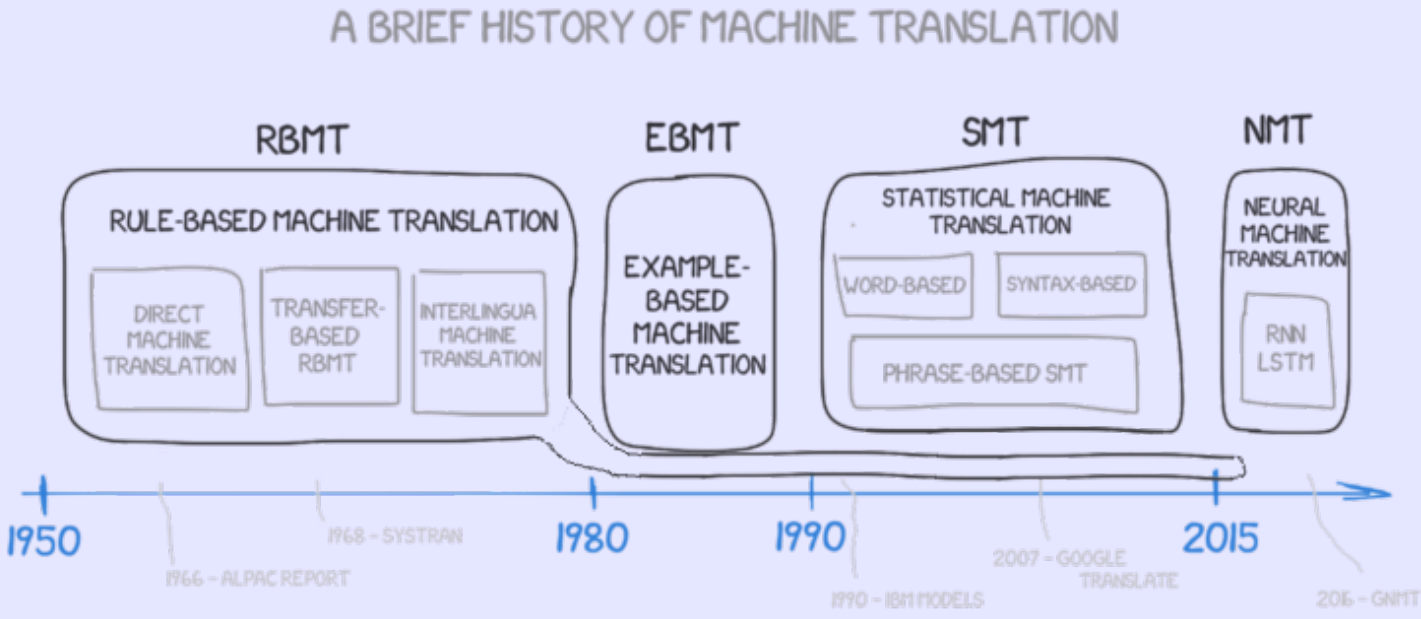
2022.05.02.

Tartalom

Tartalom

Tartalom

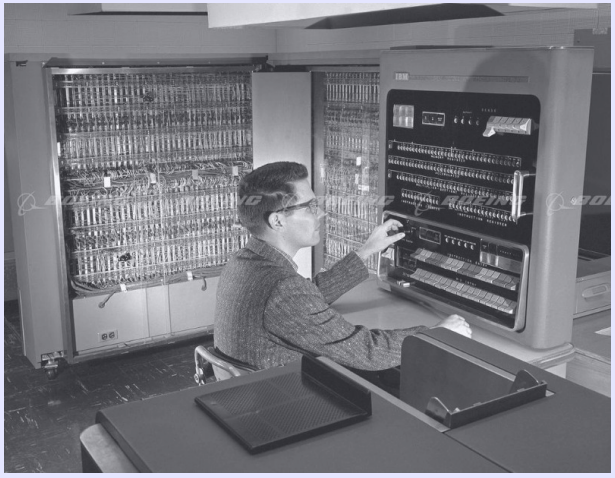
Történet a legrövidebben[1]



[1] Az ábrák forrása: `http://vas3k.com/blog/machine_translation/`

A kezdet[2]

A gép



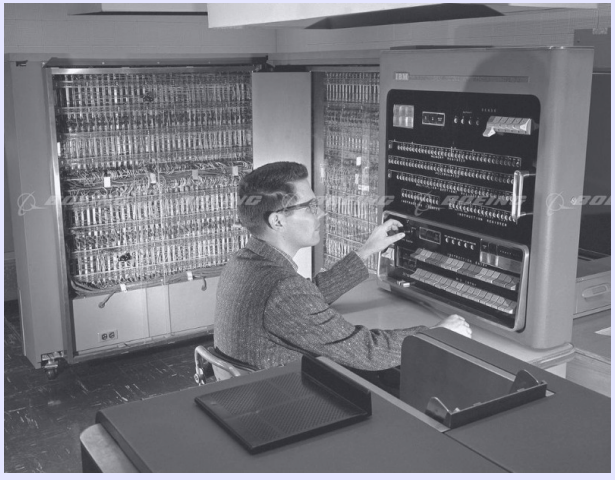
A fordítás

- 1954, IBM és New York: a számítógép képes orosz mondatokat angolra fordítani

[2] Azaz nem egészen, lásd Peter Petrovich Troyanskii-t

A kezdet[2]

A gép



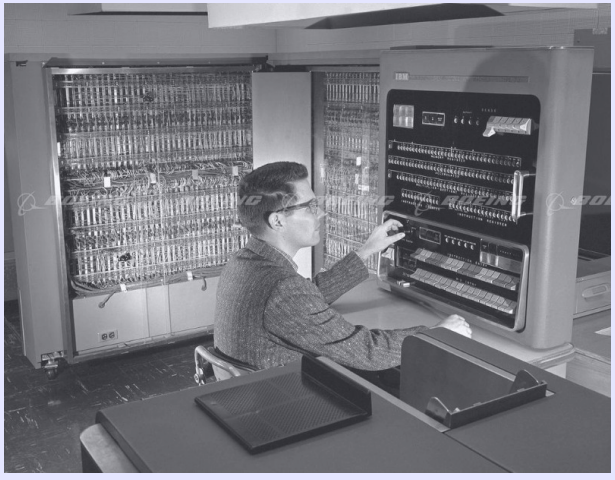
A fordítás

- 1954, IBM és New York: a számítógép képes orosz mondatokat angolra fordítani
- azaz nem egészen: gondosan válogatott példák messziről kerülve pl. a többértelműséget

[2] Azaz nem egészen, lásd Peter Petrovich Troyanskii-t

A kezdet[2]

A gép



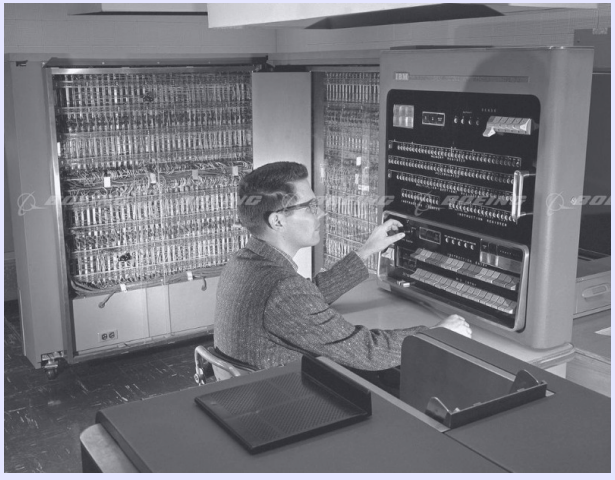
A fordítás

- 1954, IBM és New York: a számítógép képes orosz mondatokat angolra fordítani
- azaz nem egészen: gondosan válogatott példák messziről kerülve pl. a többértelműséget
- 1964, US Automatic Language Processing Advisory Committee (ALPAC): a nyár meleg, a tél hideg, nem éri meg, nem éri meg!

[2] Azaz nem egészen, lásd Peter Petrovich Troyanskii-t

A kezdet[2]

A gép



A fordítás

- 1954, IBM és New York: a számítógép képes orosz mondatokat angolra fordítani
- azaz nem egészen: gondosan válogatott példák messziről kerülve pl. a többértelműséget
- 1964, US Automatic Language Processing Advisory Committee (ALPAC): a nyár meleg, a tél hideg, nem éri meg, nem éri meg!
- 1977-2001, Kanada: METEO En→Fr időjárásjelentésfordító-rendszer

[2] Azaz nem egészen, lásd Peter Petrovich Troyanskii-t

A múlt

Mózes

MOSES

statistical
machine translation
system

A fordítás

- 2000-2016, majdnem mindenhol: statisztikai alapú gépi fordítás

A múlt

Mózes

MOSES
statistical
machine translation
system

A fordítás

- 2000-2016, majdnem mindenhol: statisztikai alapú gépi fordítás
- 2016-ig domináns (kivételek pl. PROMT, SYSTRAN, MorphoLogic)

A vég

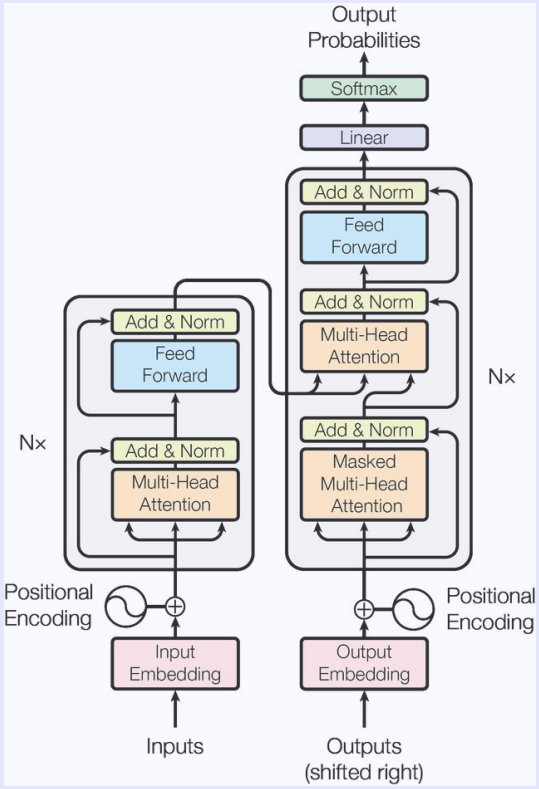
Transzformer

A fordítás

- 2000-2016, majdnem mindenhol: statisztikai alapú gépi fordítás
- 2016-ig domináns (kivételek pl. PROMT, SYSTRAN, MorphoLogic)
- 2016-, fokozatosan mindenhol: ideghálók veszik át a hatalmat

A vég

Transzformer

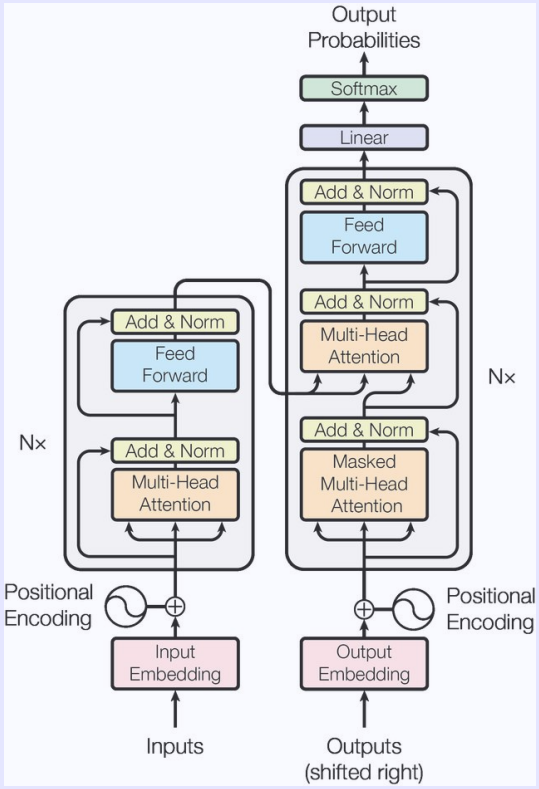


A fordítás

- 2000-2016, majdnem mindenhol: statisztikai alapú gépi fordítás
- 2016-ig domináns (kivételek pl. PROMT, SYSTRAN, MorphoLogic)
- 2016-, fokozatosan mindenhol: ideghálók veszik át a hatalmat
- 2xxx-, valahol a mátrixban: még hosszú idő (?) amíg a neurális gépi fordítás kiváltja az emberit...

A vég

Transzformer

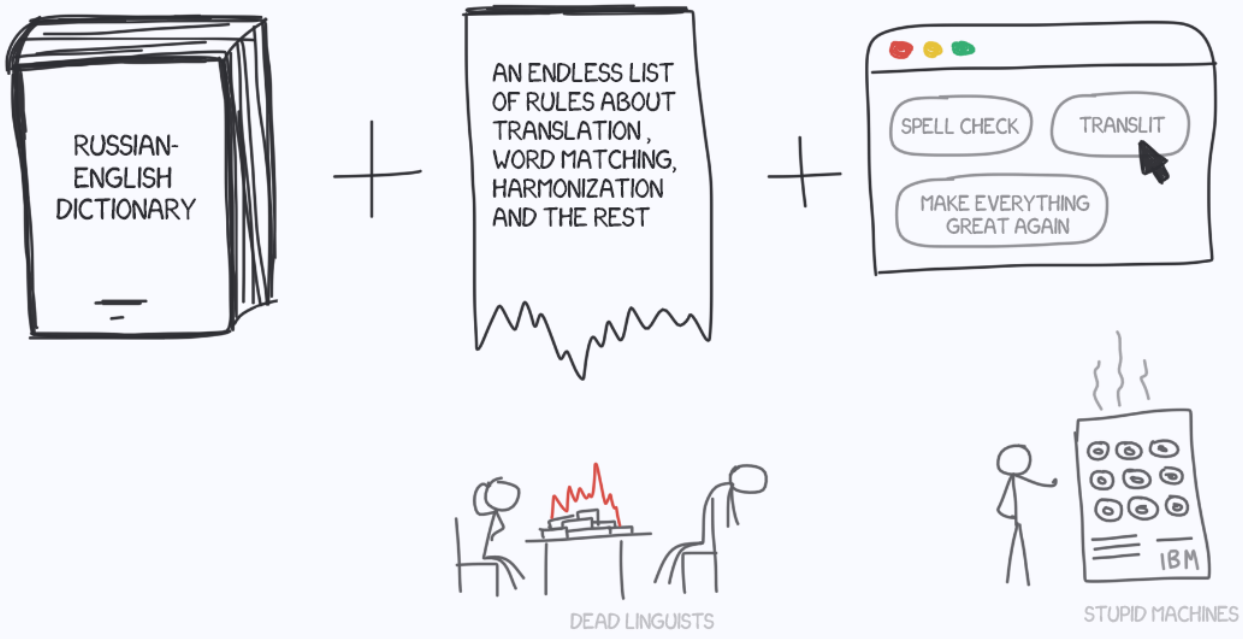


A fordítás

- 2000-2016, majdnem mindenhol: statisztikai alapú gépi fordítás
- 2016-ig domináns (kivételek pl. PROMT, SYSTRAN, MorphoLogic)
- 2016-, fokozatosan mindenhol: ideghálók veszik át a hatalmat
- 2xxx-, valahol a szingularitásban: még hosszú idő (?) amíg a neurális gépi fordítás kiváltja az emberit...

Tartalom

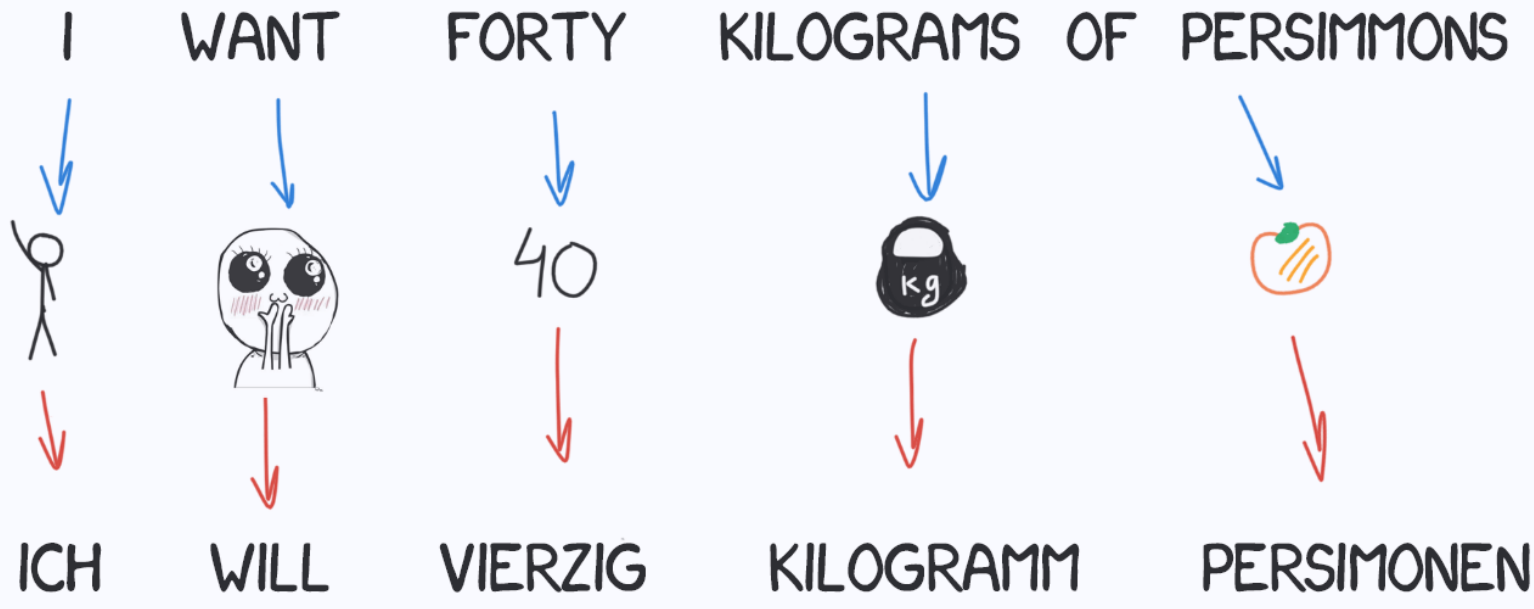
Szabályalapú gépi fordítás (RBMT)



Transzfer

```
"initiate"->"be&avat";
        $1~<syntax type="V">+direct_object=$2+modified_by_prep1=$3</syntax>;
        $2-><syntax type="N">+case=acc</syntax>;
        $3="into";
        $3~<syntax type="PREP">+direct_object=$4</syntax>;
        $4-><syntax type="N">+case=ill</syntax>;
        $3->$$$.
"insist"->"ragaszkodik";
        $1~<syntax type="V">+modified_by_prep1=$2</syntax>;
        $2="on";
        $2~<syntax type="PREP">+direct_object=$3</syntax>;
        $3-><syntax type="N">+case=all</syntax>;
        $2->$$$.
"insist"->"ragaszkodik";
        $1~<syntax type="V">+modified_by_prep1=$2</syntax>;
        $2="upon";
        $2~<syntax type="PREP">+direct_object=$3</syntax>;
        $3-><syntax type="N">+case=all</syntax>;
        $2->$$$.
```


Interlingva avagy a kakiszilva története



Vizsgakérdés

1. kérdés

Vizsgakérdés

1. kérdés

- Az interlingva alapú fordítórendszer...
  - A ugyanaz mint a transzfer alapú
  - B a tökéletes működő fordítórendszer
  - C használ(hat)ja Melcsuk és Zsolkovszkij szemantikai modelljét
  - D számítógépes nyelvi elemző eszközöket nem igényel

Vizsgakérdés

1. kérdés

Az interlingva alapú fordítórendszer...

A ugyanaz mint a transzfer alapú

B a tökéletes működő fordítórendszer

C használ(hat)ja Melcsuk és Zsolkovszkij szemantikai modelljét ✔

D számítógépes nyelvi elemző eszközöket nem igényel

Alapfogalmi kitérő

Gépi tanulás

- A (számító)gép teljesítménye egy feladat megoldásában (külső segítséggel vagy anélkül) tapasztalat alapján javul.

Alapfogalmi kitérő

Gépi tanulás

- A (számító)gép teljesítménye egy feladat megoldásában (külső segítséggel vagy anélkül) tapasztalat alapján javul.
- tapasztalat $\overset{\text{tanulás}}{\longrightarrow}$ tudás

Alapfogalmi kitérő

Gépi tanulás

- A (számító)gép teljesítménye egy feladat megoldásában (külső segítséggel vagy anélkül) tapasztalat alapján javul.
- tapasztalat $\overset{\text{tanulás}}{\longrightarrow}$ tudás
- *tapasztalat*: adat

Alapfogalmi kitérő

Gépi tanulás

- A (számító)gép teljesítménye egy feladat megoldásában (külső segítséggel vagy anélkül) tapasztalat alapján javul.
- tapasztalat $\overset{\text{tanulás}}{\longrightarrow}$ tudás
- *tapasztalat*: adat
- *tudás*: megtanult modell (függvény), melynek alapján a bemeneti (a tanulás során nem látott) adatokhoz kimeneti választ kapunk

Alapfogalmi kitérő

Gépi tanulás

- A (számító)gép teljesítménye egy feladat megoldásában (külső segítséggel vagy anélkül) tapasztalat alapján javul.
- tapasztalat $\overset{\text{tanulás}}{\longrightarrow}$ tudás
- *tapasztalat*: adat
- *tudás*: megtanult modell (függvény), melynek alapján a bemeneti (a tanulás során nem látott) adatokhoz kimeneti választ kapunk
- *reprezentáció*: az adatok (alapesetben általunk kitalált formában történő) jellemzése, formális leírása a gép számára érthető és a feladat számára hasznos módon (*feature engineering*)

Alapfogalmi kitérő

Gépi tanulás

- A (számító)gép teljesítménye egy feladat megoldásában (külső segítséggel vagy anélkül) tapasztalat alapján javul.
- tapasztalat $\overset{\text{tanulás}}{\longrightarrow}$ tudás
- *tapasztalat*: adat
- *tudás*: megtanult modell (függvény), melynek alapján a bemeneti (a tanulás során nem látott) adatokhoz kimeneti választ kapunk
- *reprezentáció*: az adatok (alapesetben általunk kitalált formában történő) jellemzése, formális leírása a gép számára érthető és a feladat számára hasznos módon (*feature engineering*)
  $\rightarrow$ nem triviális!

Alapfogalmi kitérő

Gépi tanulás

- A (számító)gép teljesítménye egy feladat megoldásában (külső segítséggel vagy anélkül) tapasztalat alapján javul.
- tapasztalat $\overset{\text{tanulás}}{\longrightarrow}$ tudás
- *tapasztalat*: adat
- *tudás*: megtanult modell (függvény), melynek alapján a bemeneti (a tanulás során nem látott) adatokhoz kimeneti választ kapunk
- *reprezentáció*: az adatok (alapesetben általunk kitalált formában történő) jellemzése, formális leírása a gép számára érthető és a feladat számára hasznos módon (*feature engineering*)
  $\rightarrow$ nem triviális!
- *tanulás*: <*bemenet*, *kimenet*> párok „végigolvasása” (felügyelt)

Vizsgakérdés

2. kérdés

Vizsgakérdés

2. kérdés

- Egy szpemszűrő modell építéséhez választott használható(nak tűnő) e-mail reprezentáció...
  - A az e-mailben található karakterek száma
  - B az átlagos szóhossz
  - C a különböző írásjelek előfordulási gyakoriságát tartalmazó n-dimenziós vektor (`['?':22, ';':  11,...]`)
  - D kiválasztott szavak előfordulását rögzítő n-dimenziós vektor (`['Dél-Afrika':1, 'bankszámla':1, 'aranybánya':1, 'örökös':1, 'csaló':0, ...]`)

Vizsgakérdés

2. kérdés

- Egy szpemszűrő modell építéséhez választott használható(nak tűnő) e-mail reprezentáció...
  - A az e-mailben található karakterek száma
  - B az átlagos szóhossz
  - C a különböző írásjelek előfordulási gyakoriságát tartalmazó n-dimenziós vektor (`['?':22, ';':  11,...]`)
  - D kiválasztott szavak előfordulását rögzítő n-dimenziós vektor (`['Dél-Afrika':1, 'bankszámla':1, 'aranybánya':1, 'örökös':1, 'csaló':0, ...]`) ✔

Statisztikai gépi fordítás (SMT)

Gépitanuljunk fordítani

Statisztikai gépi fordítás (SMT)

Tanuljunk gépifordítani

- bemenet:

Statisztikai gépi fordítás (SMT)

Tanuljunk gépifordítani

- bemenet: forrásnyelv

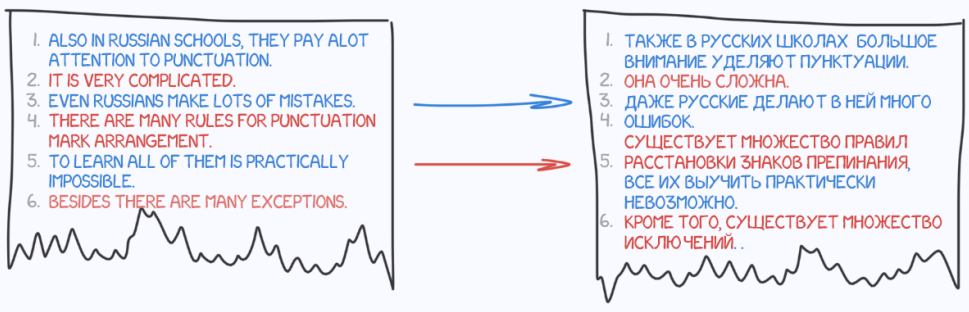
Statisztikai gépi fordítás (SMT)

Tanuljunk gépifordítani

- bemenet: forrásnyelv
- kimenet:

Statisztikai gépi fordítás (SMT)

Párhuzamos korpusz



Tanuljunk gépifordítani

- bemenet: forrásnyelv
- kimenet: célnyelv

Statisztikai gépi fordítás (SMT)

Modell

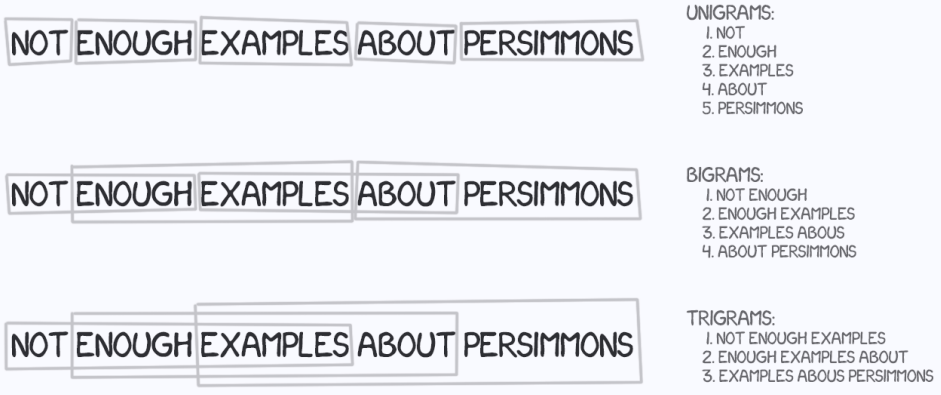
$$\hat{C} = \underset{C}{\operatorname{argmax}} P(C|F)$$

Tanuljunk gépifordítani

- bemenet: forrásnyelv
- kimenet: célnyelv
- modell: forrásnyelvi bemenethez a legvalószínűbb célnyelvi kimenet (lásd feltételes valószínűségeloszlás)

Statisztikai gépi fordítás (SMT)

Reprezentáció



Tanuljunk gépifordítani

- bemenet: forrásnyelv
- kimenet: célnyelv
- modell: forrásnyelvi bemenethez a legvalószínűbb célnyelvi kimenet (lásd feltételes valószínűségeloszlás)
- reprezentáció: pl. n-gram (n számú szó folytonos sorozata) statisztikák (meg sok más)

Evolúció – demó

Fordítsuk le

Evolúció – demó

Fordítsuk le

It is a prerequisite for enrolment in the training that there is in place a practical training agreement between the participant and a regional body of the Danish Veterinary and Food Administration (Fødevarestyrelsen).

Evolúció – demó

Fordítsuk le

It is a prerequisite for enrolment in the training that there is in place a practical training agreement between the participant and a regional body of the Danish Veterinary and Food Administration (Fødevarestyrelsen).

Szabályosan

Ez egy előfeltétel beiratkozásért bent van az edzeni azt ott helyben van egy gyakorlati tréningmegegyezés a résztvevő és a dán Veterinary és Food Administration (Fødevarestyrelsen). egy területi teste között

Evolúció – demó

Fordítsuk le

It is a prerequisite for enrolment in the training that there is in place a practical training agreement between the participant and a regional body of the Danish Veterinary and Food Administration (Fødevarestyrelsen).

Statisztikusan

A nyilvántartásba vétel előfeltétele, hogy a képzés olyan gyakorlati képzésben résztvevő közötti megállapodás egy regionális szerv és a dán állategészségügyi és élelmiszerügyi hatóságot (Fødevarestyrelsen).

Evolúció – demó

Fordítsuk le

It is a prerequisite for enrolment in the training that there is in place a practical training agreement between the participant and a regional body of the Danish Veterinary and Food Administration (Fødevarestyrelsen).

Neurálisan tegnap

A képzésen való részvétel előfeltétele, hogy gyakorlati képzési megállapodás jöjjön létre a résztvevő és a Dán Állategészségügyi és Élelmezésügyi Hivatal regionális szerve között (Fødevarestyrelsen).

Evolúció – demó

Fordítsuk le

It is a prerequisite for enrolment in the training that there is in place a practical training agreement between the participant and a regional body of the Danish Veterinary and Food Administration (Fødevarestyrelsen).

Neurálisan ma

A képzésre való beiratkozás előfeltétele, hogy a résztvevő és a dán állat-egészségügyi és élelmiszerügyi hatóság (Fødevarestyrelsen) regionális szerve között gyakorlati képzési megállapodás legyen érvényben.

Evolúció

Fordítsuk le

Evolúció

Fordítsuk le

I am looking forward to continue our cooperation to ensure consistency of EU external policies.

Evolúció

Fordítsuk le

I am looking forward to continue our cooperation to ensure consistency of EU external policies.

Szabályosan

Haladónak tűnök hogy folytassam az együttműködésünket hogy biztosítsam az EU külső politikák egyenletességét.

Evolúció

Fordítsuk le

I am looking forward to continue our cooperation to ensure consistency of EU external policies.

Statisztikusan

Örömmel várom, hogy folytatjuk együttműködésünket annak érdekében, hogy azok összhangban álljanak az Unió külső politikáit.

Evolúció

Fordítsuk le

I am looking forward to continue our cooperation to ensure consistency of EU external policies.

Neurálisan tegnap

Várakozással tekintek együttműködésünk folytatása elé az EU külső politikái közötti összhang biztosítása érdekében.

Evolúció

Fordítsuk le

I am looking forward to continue our cooperation to ensure consistency of EU external policies.

Neurálisan ma

Várakozással tekintek az együttműködés folytatása elé annak érdekében, hogy biztosítsuk az EU külső politikáinak következetességét.

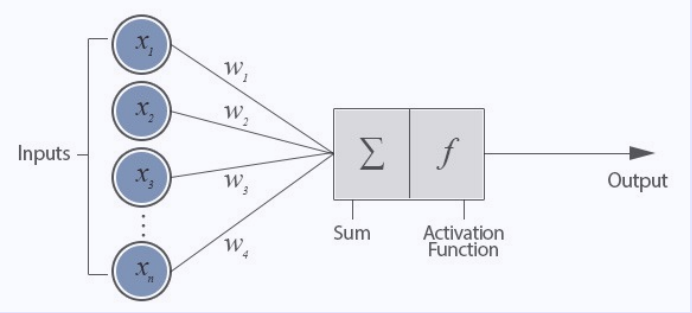
Tartalom

(Mesterséges) neurális háló

Építőelemek

(Mesterséges) neurális háló

Neuron

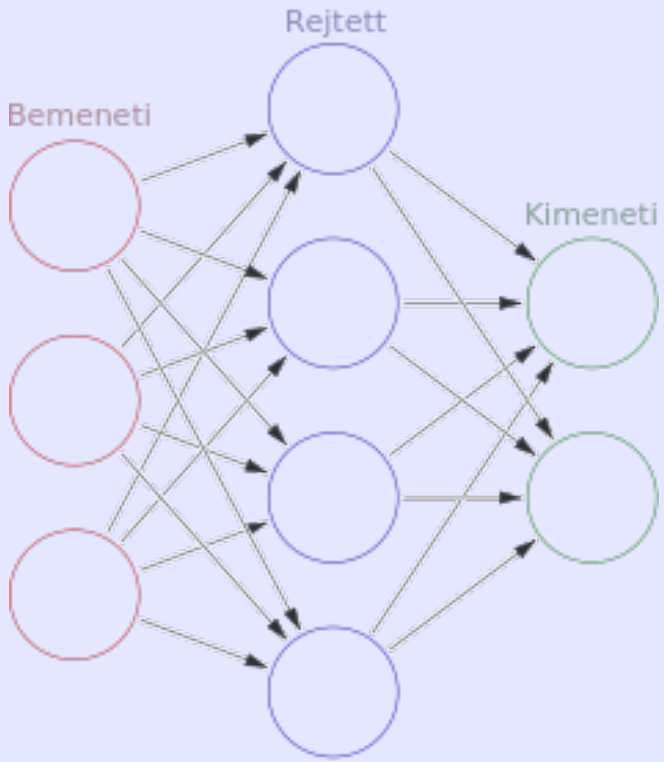


Építőelemek

- neuron: a bemenetek súlyozott összegét képzi, majd erre még alkalmaz egy függvényt...

(Mesterséges) neurális háló

Rétegek

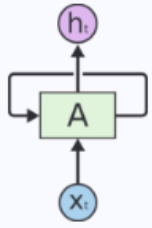


Építőelemek

- neuron: a bemenetek súlyozott összegét képzi, majd erre még alkalmaz egy függvényt...
- rétegek: hasonló típusú neuronok egy csoportja (bemeneti (input layer), rejtett (hidden layer(s)), kimeneti (ouput layer))

(Mesterséges) neurális háló

Rekurrens neurális háló



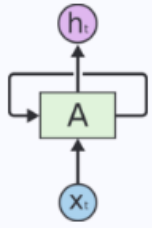
Recurrent Neural Networks have loops.

Építőelemek

- neuron: a bemenetek súlyozott összegét képzi, majd erre még alkalmaz egy függvényt...
- rétegek: hasonló típusú neuronok egy csoportja (bemeneti (input layer), rejtett (hidden layer(s)), kimeneti (ouput layer))
- rekurrens hálózat: megelőző bement(ek)re emlékező visszacsatolást tartalmaz $\rightarrow$ változó hosszúságú bemenetek kezelése (pl. .....?)

(Mesterséges) neurális háló

Rekurrens neurális háló



Recurrent Neural Networks have loops.

Építőelemek

- neuron: a bemenetek súlyozott összegét képzi, majd erre még alkalmaz egy függvényt...
- rétegek: hasonló típusú neuronok egy csoportja (bemeneti (input layer), rejtett (hidden layer(s)), kimeneti (ouput layer))
- rekurrens hálózat: megelőző bement(ek)re emlékező visszacsatolást tartalmaz $\rightarrow$ változó hosszúságú bemenetek kezelése (pl. mondat)

Alapfogalmi kitérő

Reprezentációs tanulás

Alapfogalmi kitérő

Reprezentációs tanulás

Nemcsak a bemenet-kimenet leképezést tanuljuk meg (automatikusan), hanem az adatnak az adott feladathoz optimális jellemzőit is (v.ö. *manual feature engineering*).

Mélytanulás – *deep learning*

Alapfogalmi kitérő

Reprezentációs tanulás

Nemcsak a bemenet-kimenet leképezést tanuljuk meg (automatikusan), hanem az adatnak az adott feladathoz optimális jellemzőit is (v.ö. *manual feature engineering*).

Mélytanulás – *deep learning*

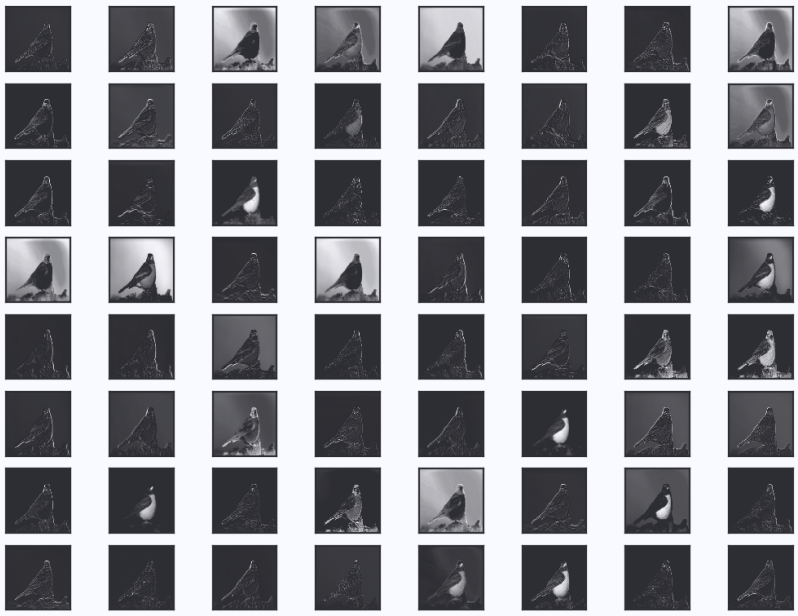
A betanult model több egymásra épülő rétegből áll, melyek az adat (egyre absztraktabb) reprezentációját hordozzák.
Pl. többrétegú neurális háló

Alapfogalmi kitérő

Madarak

Alapfogalmi kitérő

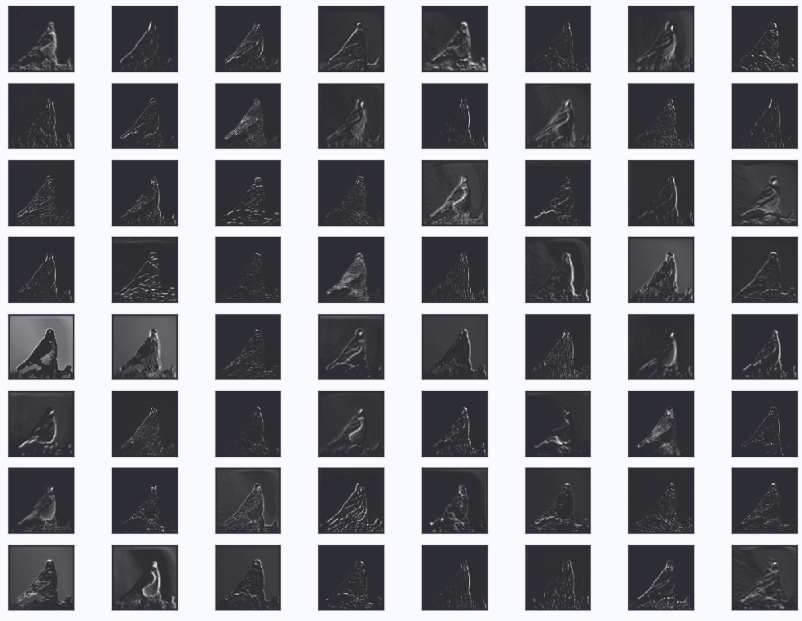
Mélytanult madarak[3]



[3] https://machinelearningmastery.com/how-to-visualize-filters-and-feature-maps-in-convolutional-neural-networks/

Alapfogalmi kitérő

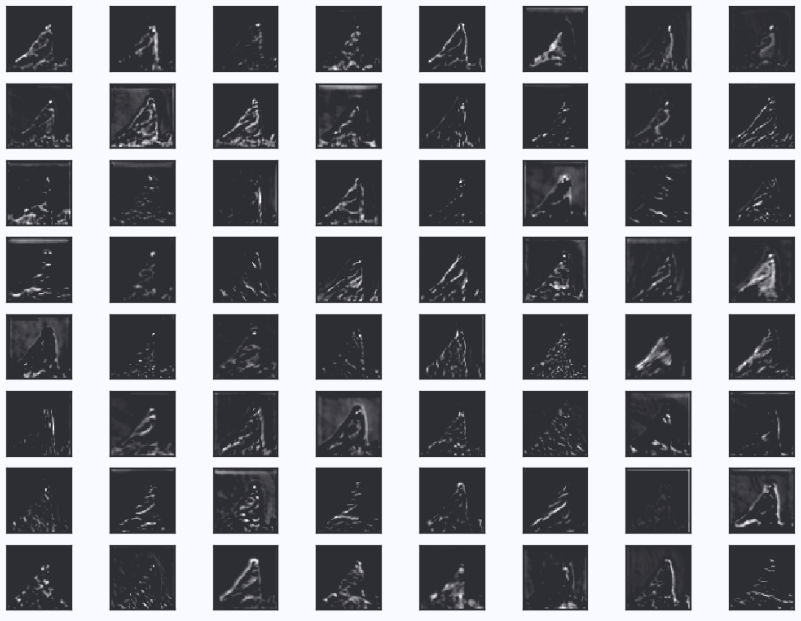
Mélytanult madarak[3]



[3] https://machinelearningmastery.com/how-to-visualize-filters-and-feature-maps-in-convolutional-neural-networks/

Alapfogalmi kitérő

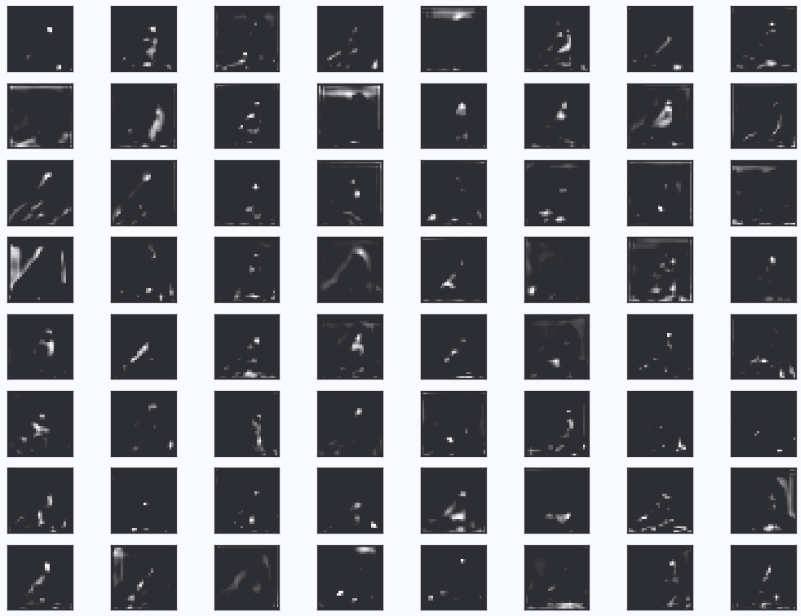
Mélytanult madarak[3]



[3] https://machinelearningmastery.com/how-to-visualize-filters-and-feature-maps-in-convolutional-neural-networks/

Alapfogalmi kitérő

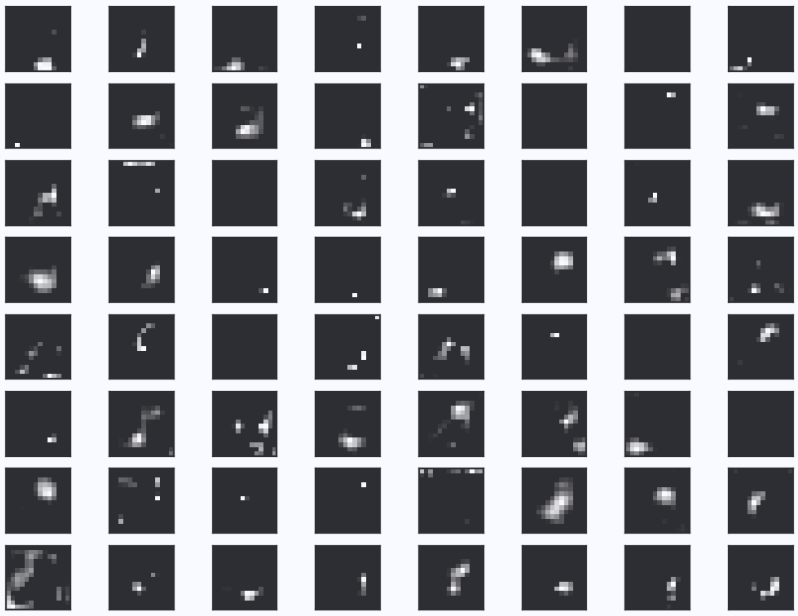
Mélytanult madarak[3]



[3] https://machinelearningmastery.com/how-to-visualize-filters-and-feature-maps-in-convolutional-neural-networks/

Alapfogalmi kitérő

Mélytanult madarak[3]

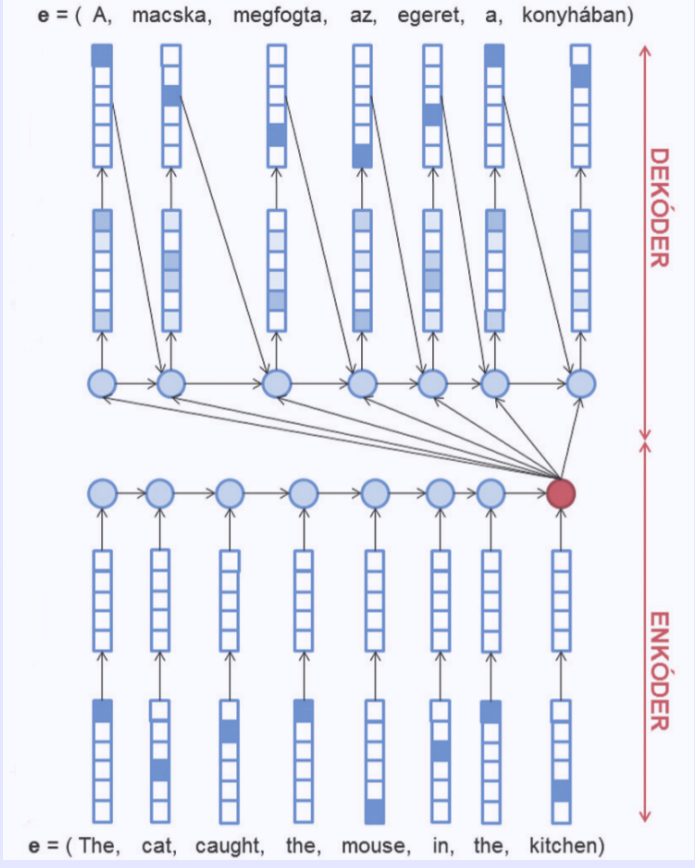


[3] https://machinelearningmastery.com/how-to-visualize-filters-and-feature-maps-in-convolutional-neural-networks/

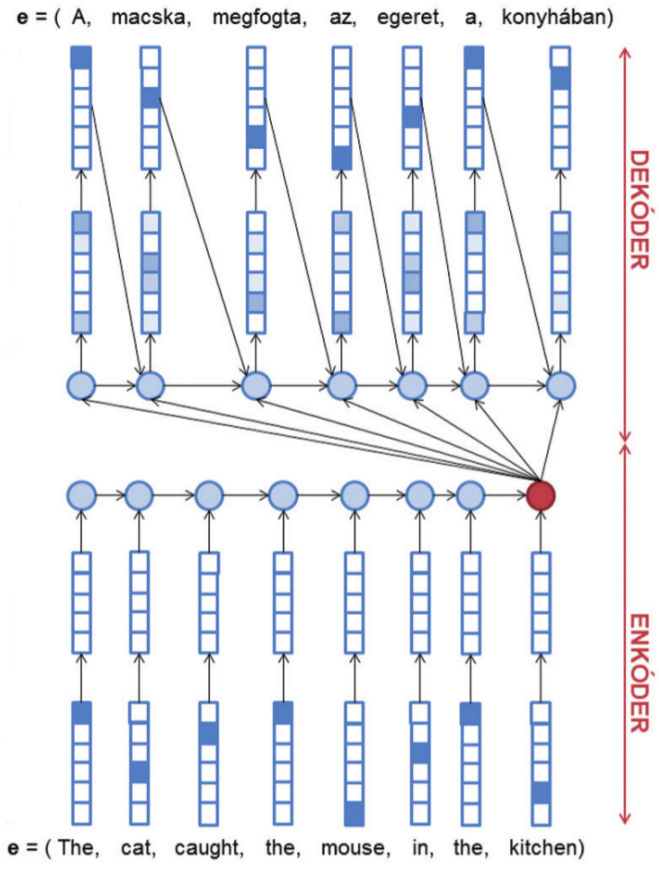
Neurális kódoló és dekódoló

Neurális kódoló és dekódoló

Órarugó gerincű lesből támadó gépi mélytanuló neurális fordító

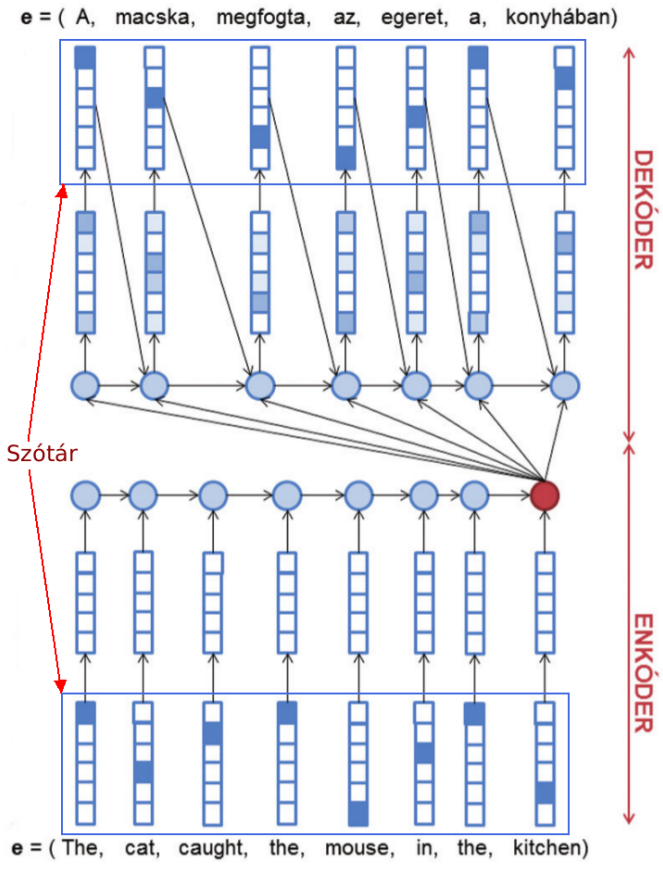


Neurális kódoló és dekódoló



Felépítés és működés

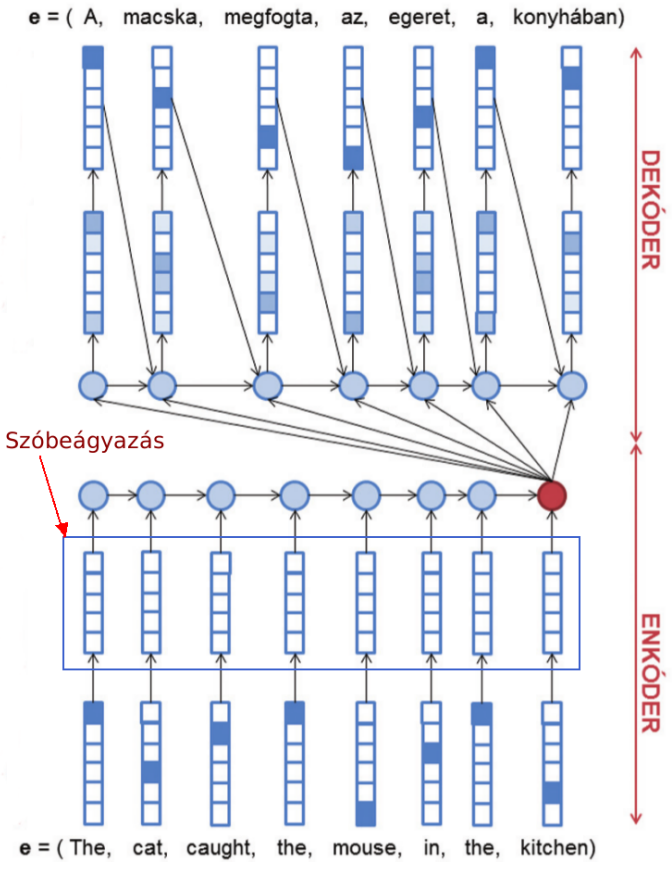
Neurális kódoló és dekódoló



Felépítés és működés

- szótár: 1-K kódolás, $|V|$ (szótárméret) hosszú vektor `[0,...,0,1,0,...,0]`

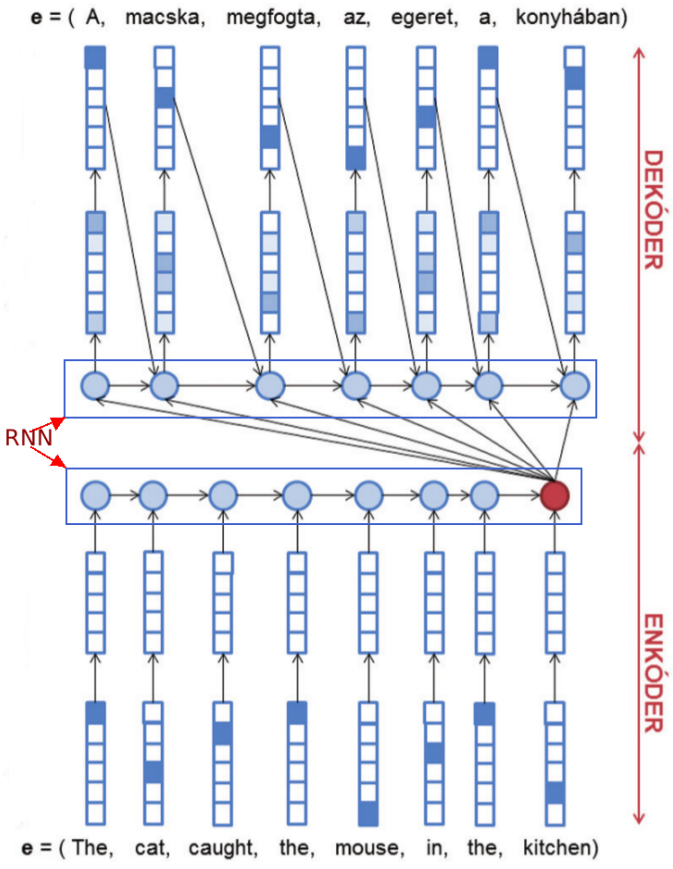
Neurális kódoló és dekódoló



Felépítés és működés

- szótár: 1-K kódolás, $|V|$ (szótárméret) hosszú vektor `[0,...,0,1,0,...,0]`
- szóbeágyazás: szóvektor $\times$ beágyazási mátrix ($|V|$ oszlop, kb. 500 sor)

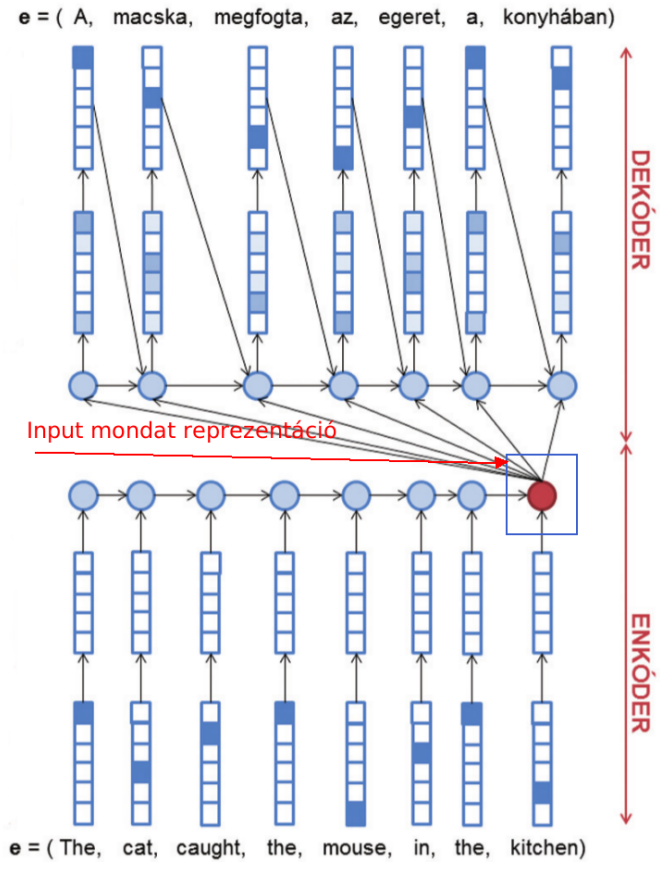
Neurális kódoló és dekódoló



Felépítés és működés

- szótár: 1-K kódolás, $|V|$ (szótárméret) hosszú vektor `[0,...,0,1,0,...,0]`
- szóbeágyazás: szóvektor $\times$ beágyazási mátrix ($|V|$ oszlop, kb. 500 sor)
- rekurrens háló (egy vagy több rétegű)

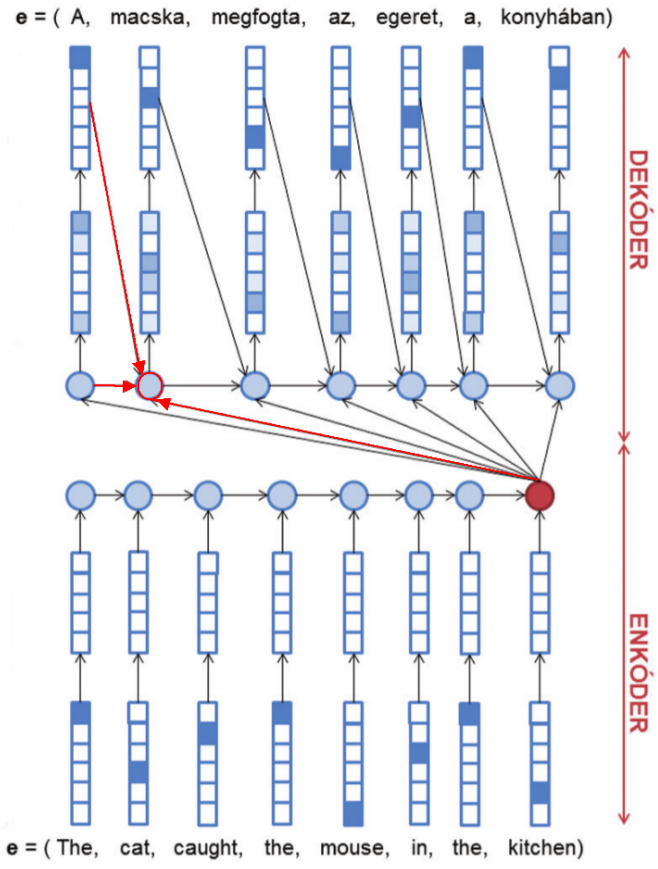
Neurális kódoló és dekódoló



Felépítés és működés

- szótár: 1-K kódolás, $|V|$ (szótárméret) hosszú vektor `[0,...,0,1,0,...,0]`
- szóbeágyazás: szóvektor $\times$ beágyazási mátrix ($|V|$ oszlop, kb. 500 sor)
- rekurrens háló (egy vagy több rétegű)
- forrásnyelvi mondatot reprezentáló vektor

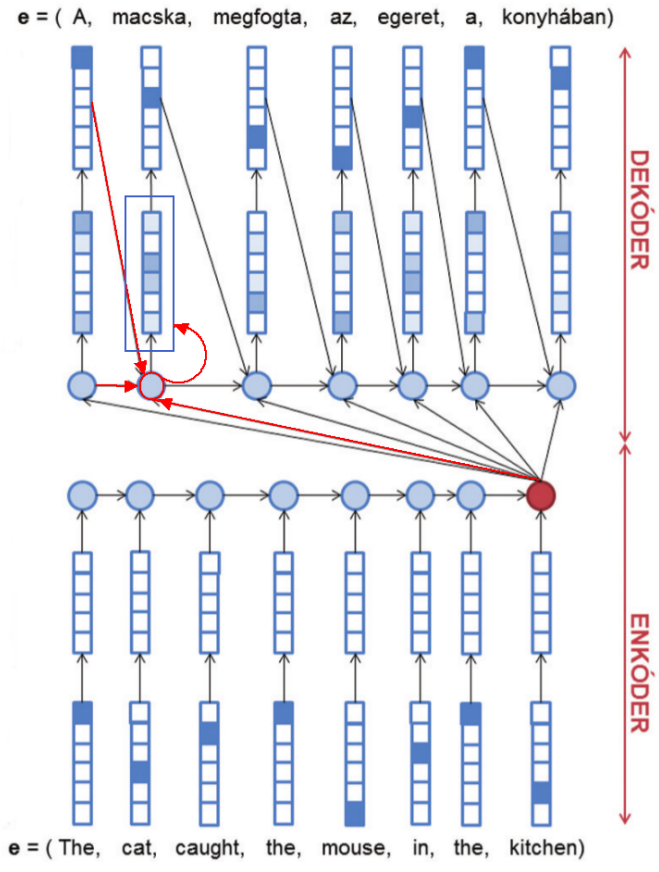
Neurális kódoló és dekódoló



Felépítés és működés

- szótár: 1-K kódolás, $|V|$ (szótárméret) hosszú vektor `[0,...,0,1,0,...,0]`
- szóbeágyazás: szóvektor $\times$ beágyazási mátrix ($|V|$ oszlop, kb. 500 sor)
- rekurrens háló (egy vagy több rétegű)
- forrásnyelvi mondatot reprezentáló vektor
- dekóder állapot: $<$*input reprezentáció, megelőző állapot, megelőző célnyelvi szó*$>$

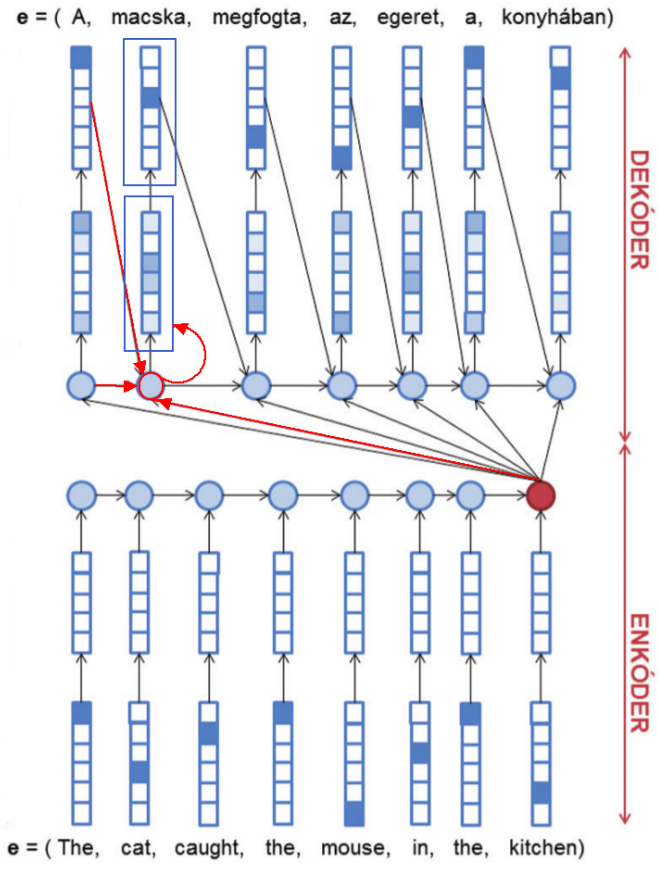
Neurális kódoló és dekódoló



Felépítés és működés

- szótár: 1-K kódolás, $|V|$ (szótárméret) hosszú vektor $[0,\ldots,0,1,0,\ldots,0]$
- szóbeágyazás: szóvektor $\times$ beágyazási mátrix ($|V|$ oszlop, kb. 500 sor)
- rekurrens háló (egy vagy több rétegű)
- forrásnyelvi mondatot reprezentáló vektor
- dekóder állapot: $<$*input reprezentáció, megelőző állapot, megelőző célnyelvi szó*$>$
- célnyelvi szavak eloszlása: az adott állapotban melyik szó milyen valószínű

Neurális kódoló és dekódoló



Felépítés és működés

- szótár: 1-K kódolás, $|V|$ (szótárméret) hosszú vektor $[0,\ldots,0,1,0,\ldots,0]$
- szóbeágyazás: szóvektor $\times$ beágyazási mátrix ($|V|$ oszlop, kb. 500 sor)
- rekurrens háló (egy vagy több rétegű)
- forrásnyelvi mondatot reprezentáló vektor
- dekóder állapot: $<$*input reprezentáció, megelőző állapot, megelőző célnyelvi szó*$>$
- célnyelvi szavak eloszlása: az adott állapotban melyik szó milyen valószínű

Várbeágyazás[4]

[4]Novák et al. [2018]

Mitől lesz jó a célnyelvi mondat

A cél

- (tegyük fel, hogy készen kapunk egy modellt)
- még mindig: $\hat{C} = \text{argmax}_C P(C|F)$

Mitől lesz jó a célnyelvi mondat

A cél

- (tegyük fel, hogy készen kapunk egy modellt)
- még mindig: $\hat{C} = \text{argmax}_C P(C|F)$

A módszer

- ?

Vizsgakérdés

3. kérdés

Vizsgakérdés

3. kérdés

- Hogyan találjuk meg az optimális célnyelvi mondatot?

Vizsgakérdés

3. kérdés

- Hogyan találjuk meg az optimális célnyelvi mondatot?
  - A számoljuk ki minden célnyelvi mondat valószínűségét a modellünk szerint és válasszuk a legvalószínűbbet

Vizsgakérdés

3. kérdés

- Hogyan találjuk meg az optimális célnyelvi mondatot?
  - A számoljuk ki minden célnyelvi mondat valószínűségét a modellünk szerint és válasszuk a legvalószínűbbet ✘ ($|V|^N$)

Vizsgakérdés

3. kérdés

- Hogyan találjuk meg az optimális célnyelvi mondatot?
  - A számoljuk ki minden célnyelvi mondat valószínűségét a modellünk szerint és válasszuk a legvalószínűbbet ✘ ($|V|^N$)
  - B minden dekóder állapotban válasszuk véletlenszerűen a célszót

Vizsgakérdés

3. kérdés

- Hogyan találjuk meg az optimális célnyelvi mondatot?
  - A számoljuk ki minden célnyelvi mondat valószínűségét a modellünk szerint és válasszuk a legvalószínűbbet ✘ ($|V|^N$)
  - B minden dekóder állapotban válasszuk véletlenszerűen a célszót ✘

Vizsgakérdés

3. kérdés

- Hogyan találjuk meg az optimális célnyelvi mondatot?
  - A számoljuk ki minden célnyelvi mondat valószínűségét a modellünk szerint és válasszuk a legvalószínűbbet ✘ ($|V|^N$)
  - B minden dekóder állapotban válasszuk véletlenszerűen a célszót ✘
  - C minden dekóder állapotban válasszuk a legvalószínűbb célszót

Vizsgakérdés

3. kérdés

- Hogyan találjuk meg az optimális célnyelvi mondatot?
  - A számoljuk ki minden célnyelvi mondat valószínűségét a modellünk szerint és válasszuk a legvalószínűbbet ✘ ($|V|^N$)
  - B minden dekóder állapotban válasszuk véletlenszerűen a célszót ✘
  - C minden dekóder állapotban válasszuk a legvalószínűbb célszót ✘

Vizsgakérdés

3. kérdés

- Hogyan találjuk meg az optimális célnyelvi mondatot?
  - A számoljuk ki minden célnyelvi mondat valószínűségét a modellünk szerint és válasszuk a legvalószínűbbet ✘ ($|V|^N$)
  - B minden dekóder állapotban válasszuk véletlenszerűen a célszót ✘
  - C minden dekóder állapotban válasszuk a legvalószínűbb célszót ✘
  - D minden lépésben csak maximum K számú legvalószínűbb szósorral számolunk (K-szélességű nyalábolt keresés)

Vizsgakérdés

3. kérdés

- Hogyan találjuk meg az optimális célnyelvi mondatot?
  - A számoljuk ki minden célnyelvi mondat valószínűségét a modellünk szerint és válasszuk a legvalószínűbbet ✘ ($|V|^N$)
  - B minden dekóder állapotban válasszuk véletlenszerűen a célszót ✘
  - C minden dekóder állapotban válasszuk a legvalószínűbb célszót ✘
  - D minden lépésben csak maximum K számú legvalószínűbb szósorral számolunk (K-szélességű nyalábolt keresés) ✔ ($2 < K < 10$)

Mitől lesz jó a célnyelvi mondat

A cél

- (tegyük fel, hogy készen kapunk egy modellt)
- még mindig: $\hat{C} = \text{argmax}_C P(C|F)$

A módszer

- bímszörcs

Jó modellre van szükség $\rightarrow$ tanítás

- tanító adat: párhuzamos korpusz $<$*forrás, cél*$>$ mondatpárjai
- a modell paramétereit (a számokat a mátrixokban) addig változtatjuk amíg a forrásmondatok modell által generált fordítása elég hasonló nem lesz a tanító adat célmondataihoz

Haladóbb modell

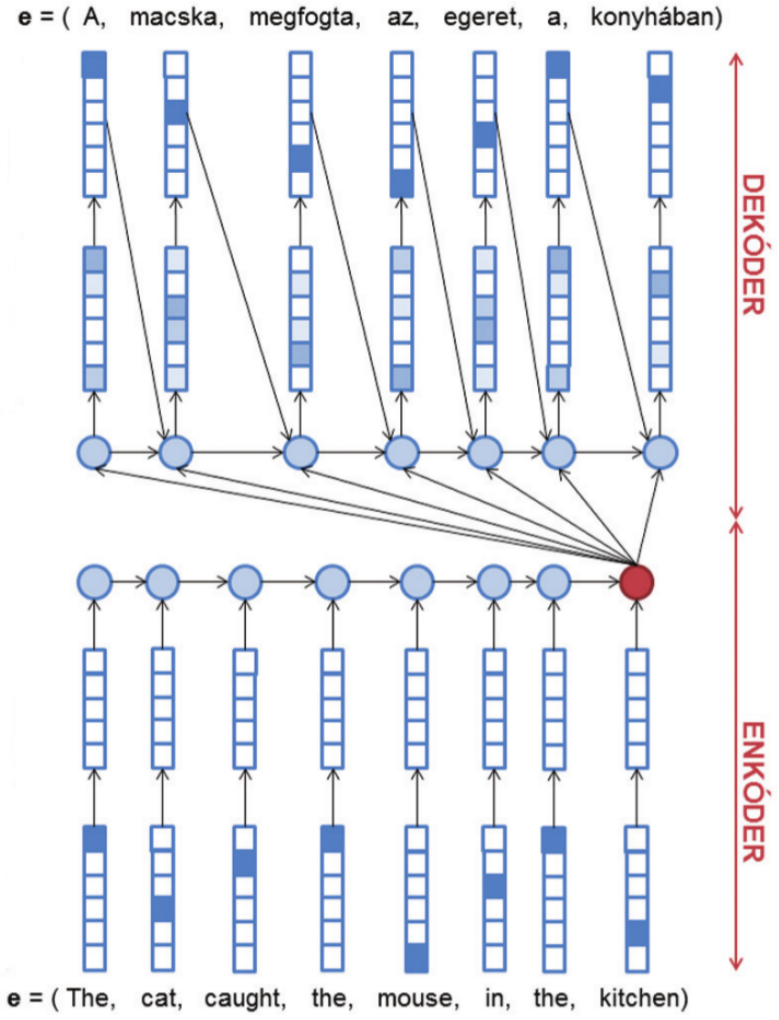
A vanília RNN problémája

- hosszabb mondat $\rightarrow$ rosszabb fordítás
- mindenkinek ez (lesz) a baja, de itt nagyon súlyos

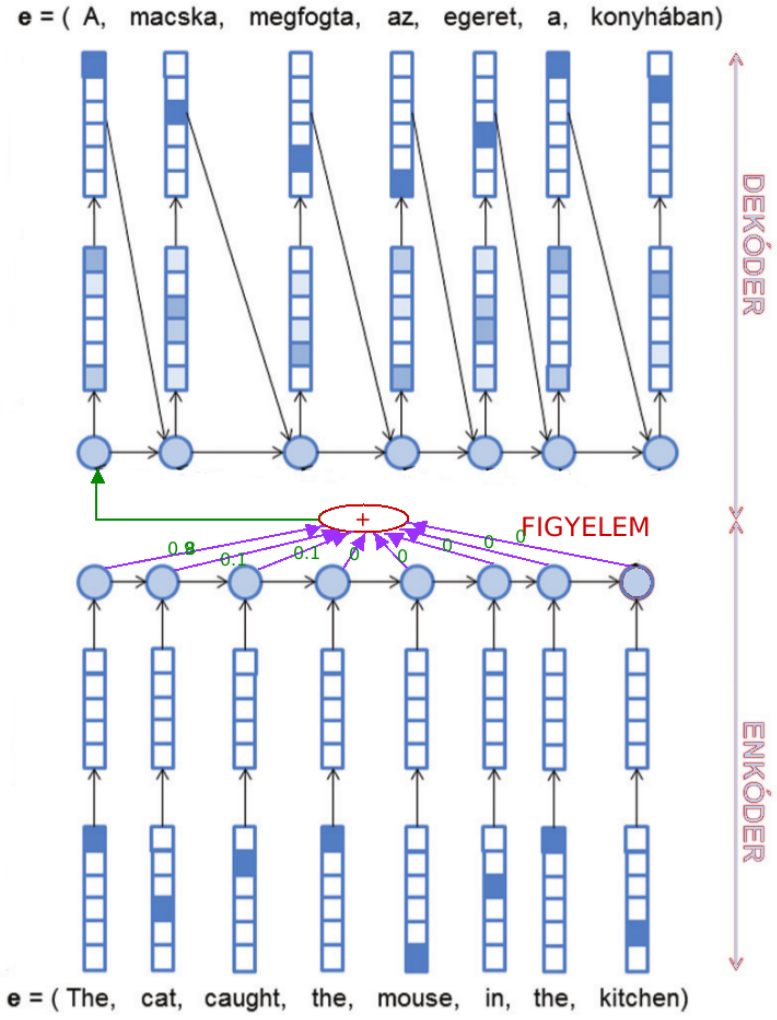
Megoldás

- figyeljünk arra, a forrásmondat melyik részét mikor fordítjuk
  $\rightarrow$ A Figyelmes RNN (RNN with soft attention)

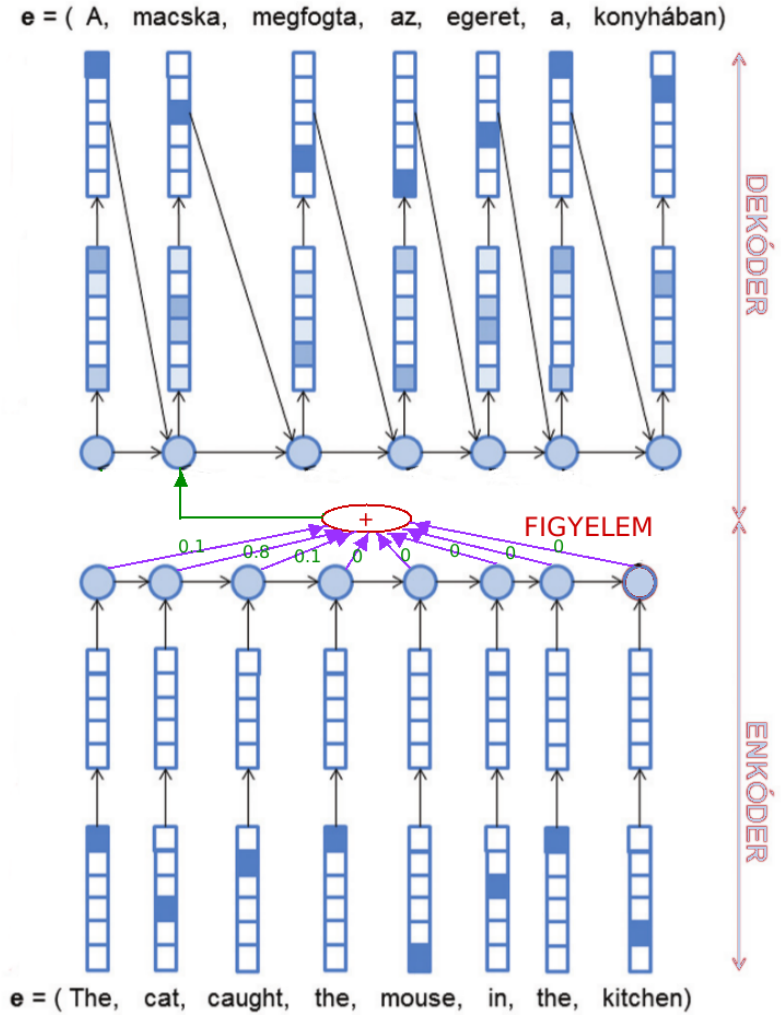
Vanília RNN


e = (A, macska, megfogta, az, egeret, a, konyhában)
DEKÓDER
ENKÓDER
e = (The, cat, caught, the, mouse, in, the, kitchen)

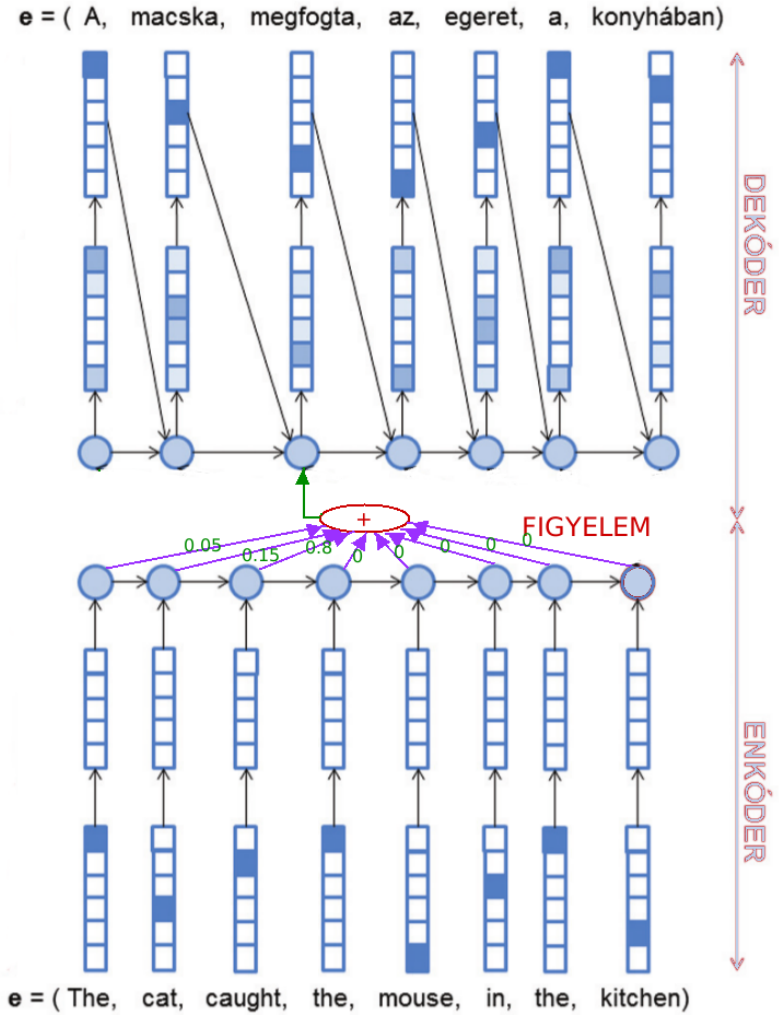
Figyelmes RNN



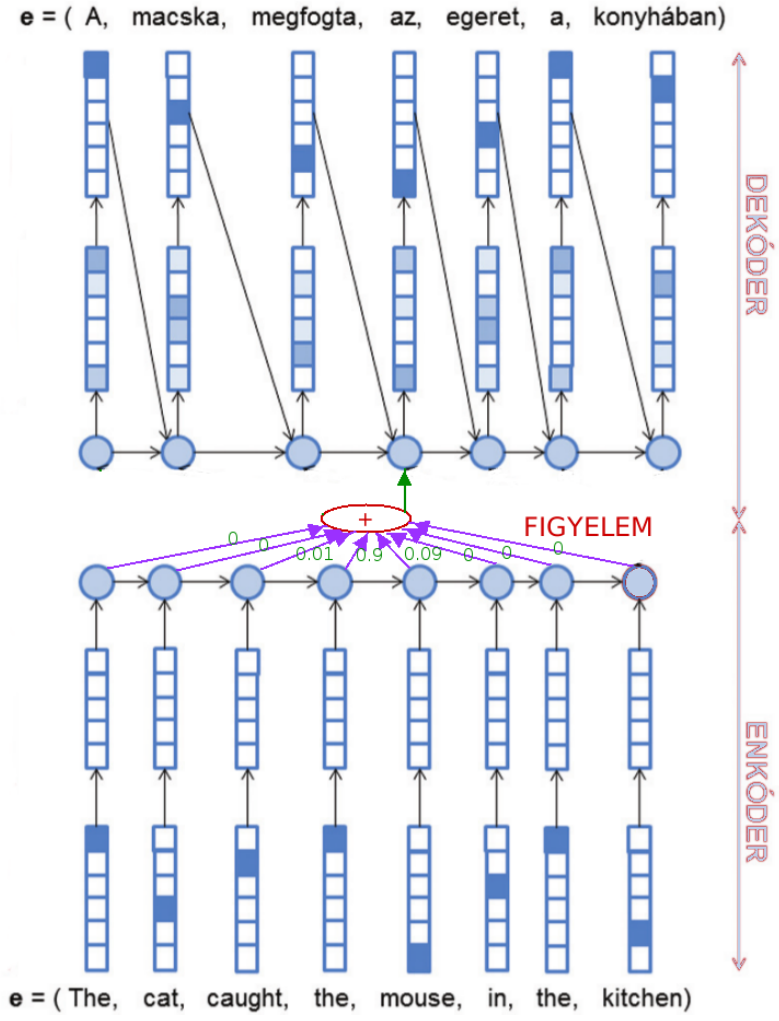
Figyelmes RNN



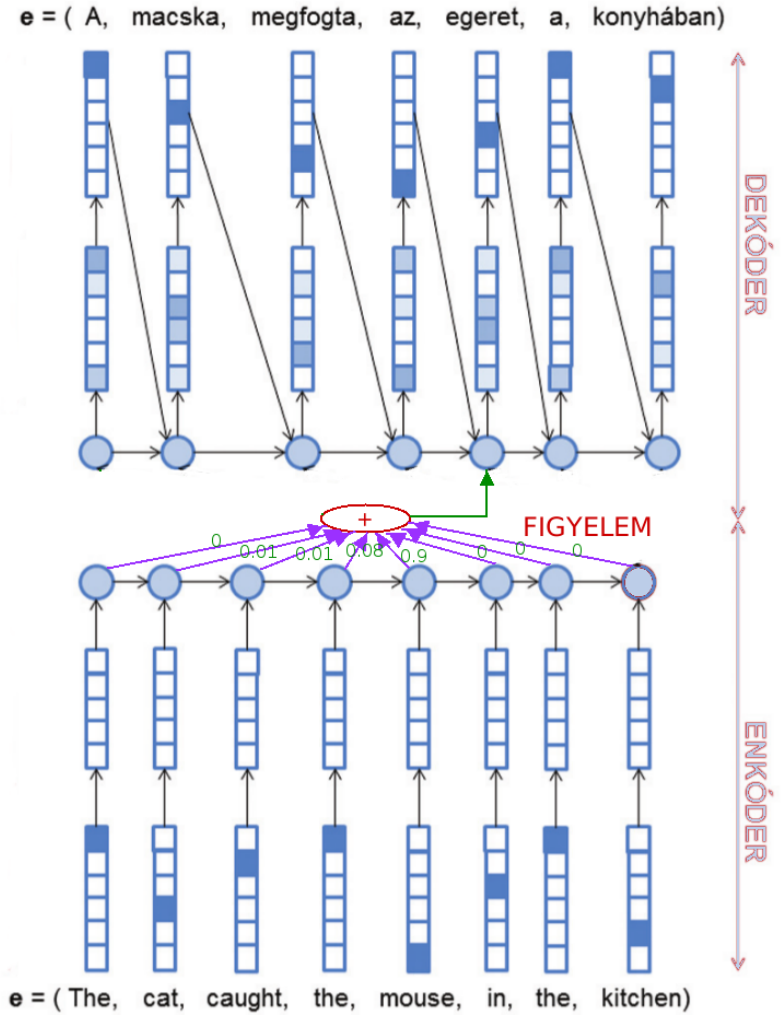
Figyelmes RNN



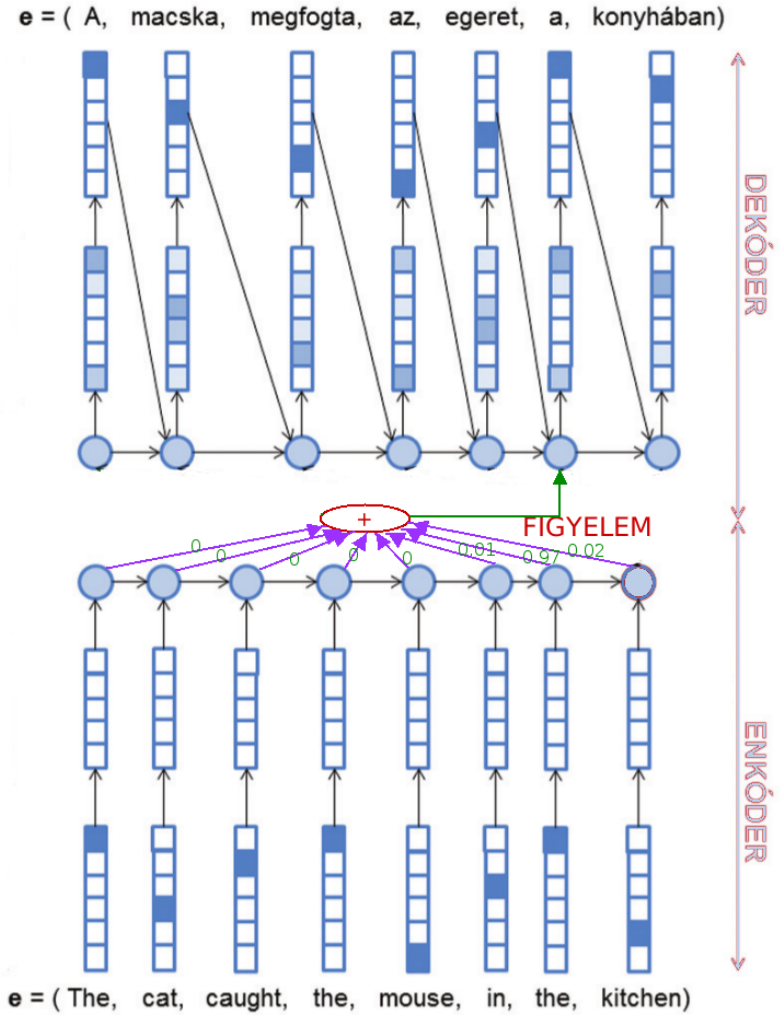
Figyelmes RNN



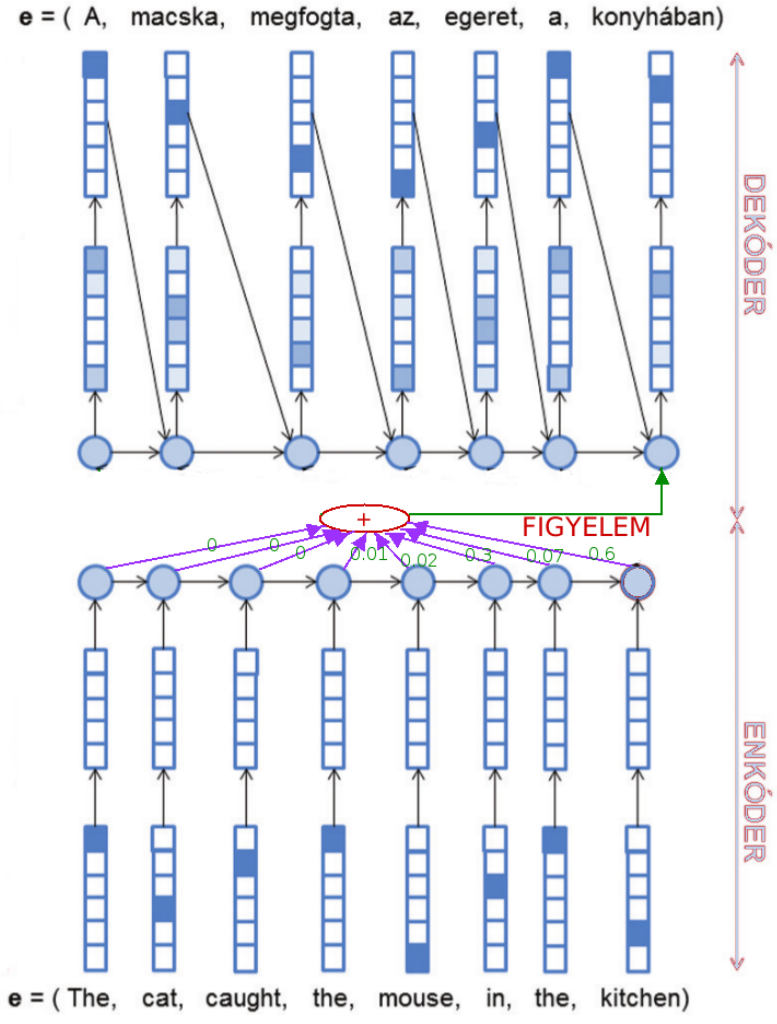
Figyelmes RNN



Figyelmes RNN



Figyelmes RNN



Figyelem, figyelem

NMT >> SMT (PBMT), RBMT,...

Miért (sokkal) jobb

NMT >> SMT (PBMT), RBMT,...

Miért (sokkal) jobb

- globális kontextus (egész mondat reprezentációja)

NMT >> SMT (PBMT), RBMT,...

Miért (sokkal) jobb

- globális kontextus (egész mondat reprezentációja)
- globális optimalizálás

NMT >> SMT (PBMT), RBMT,...

Miért (sokkal) jobb

- globális kontextus (egész mondat reprezentációja)
- globális optimalizálás
- kisméretű modell

NMT >> SMT (PBMT), RBMT,...

Miért (sokkal) jobb

- globális kontextus (egész mondat reprezentációja)
- globális optimalizálás
- kisméretű modell

De ...

NMT >> SMT (PBMT), RBMT,...

Miért (sokkal) jobb

- globális kontextus (egész mondat reprezentációja)
- globális optimalizálás
- kisméretű modell

De ...

- fekete doboz

NMT >> SMT (PBMT), RBMT,...

Miért (sokkal) jobb

- globális kontextus (egész mondat reprezentációja)
- globális optimalizálás
- kisméretű modell

De ...

- fekete doboz
- korlátozott szótárméret

NMT >> SMT (PBMT), RBMT,...

Miért (sokkal) jobb

- globális kontextus (egész mondat reprezentációja)
- globális optimalizálás
- kisméretű modell

De ...

- fekete doboz
- korlátozott szótárméret
- (bőbeszédű) túlfordítás

NMT >> SMT (PBMT), RBMT,...

Miért (sokkal) jobb

- globális kontextus (egész mondat reprezentációja)
- globális optimalizálás
- kisméretű modell

De ...

- fekete doboz
- korlátozott szótárméret
- (bőbeszédű) túlfordítás
- alulfordítás (hiányzó forrásnyelvi elemek)

NMT >> SMT (PBMT), RBMT,...

Miért (sokkal) jobb

- globális kontextus (egész mondat reprezentációja)
- globális optimalizálás
- kisméretű modell

De ...

- fekete doboz
- korlátozott szótárméret
- (bőbeszédű) túlfordítás
- alulfordítás (hiányzó forrásnyelvi elemek)
- jólformáltság $\leftrightarrow$ adekvátság

NMT >> SMT (PBMT), RBMT,...

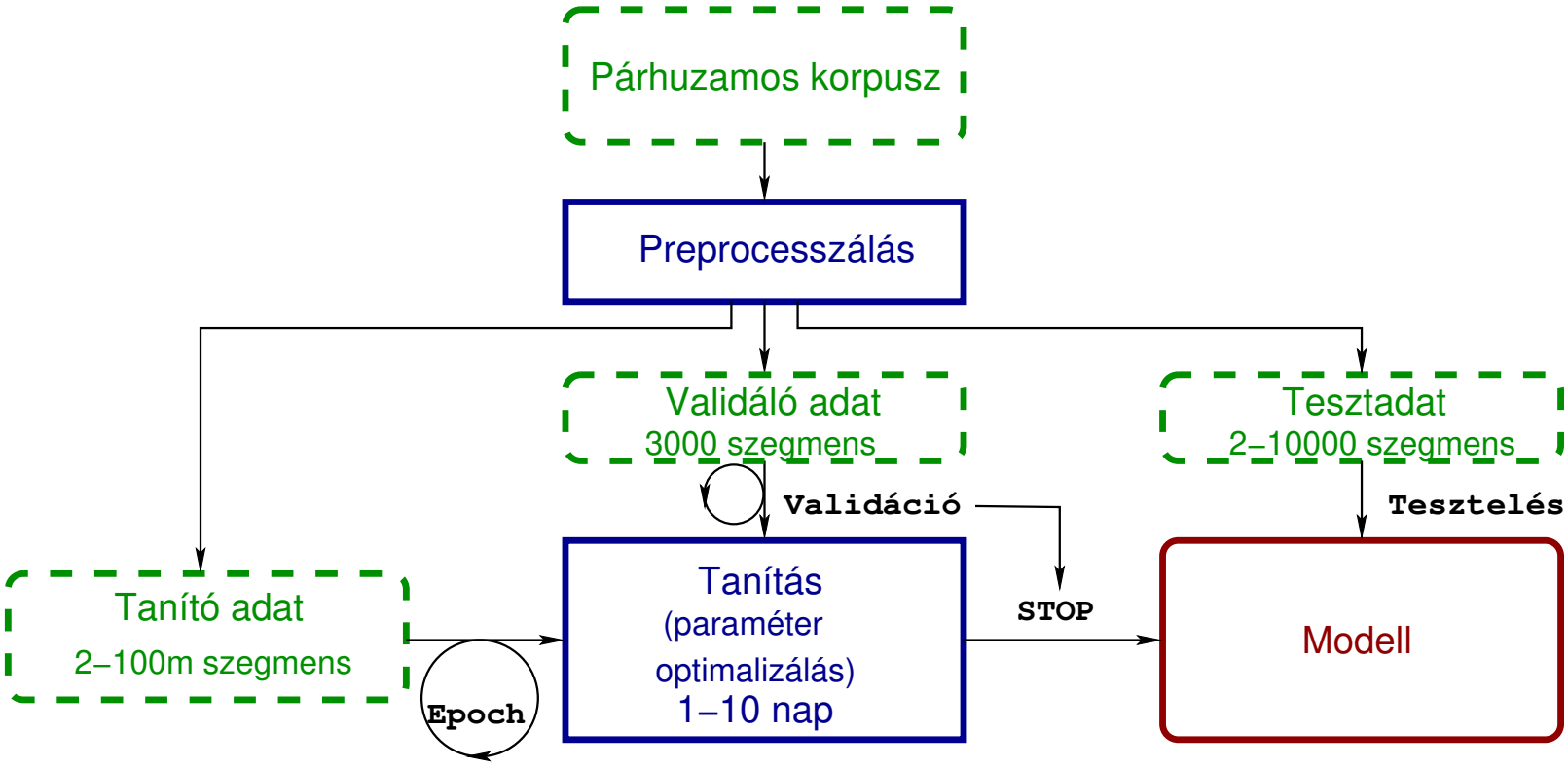
Miért (sokkal) jobb

- globális kontextus (egész mondat reprezentációja)
- globális optimalizálás
- kisméretű modell

De ...

- fekete doboz
- korlátozott szótárméret
- (bőbeszédű) túlfordítás
- alulfordítás (hiányzó forrásnyelvi elemek)
- jólformáltság $\leftrightarrow$ adekvátság
- zajérzékenység

Hogyan készül?



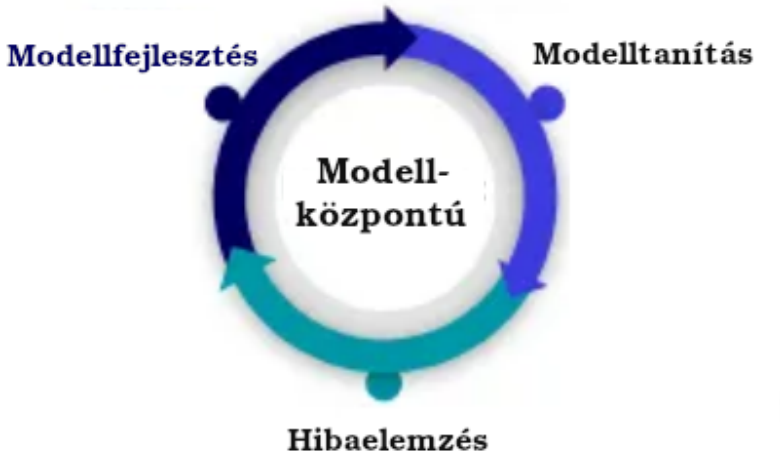
Tartalom

Trendek

"Everyone wants to do the model work, not the data work"[a]

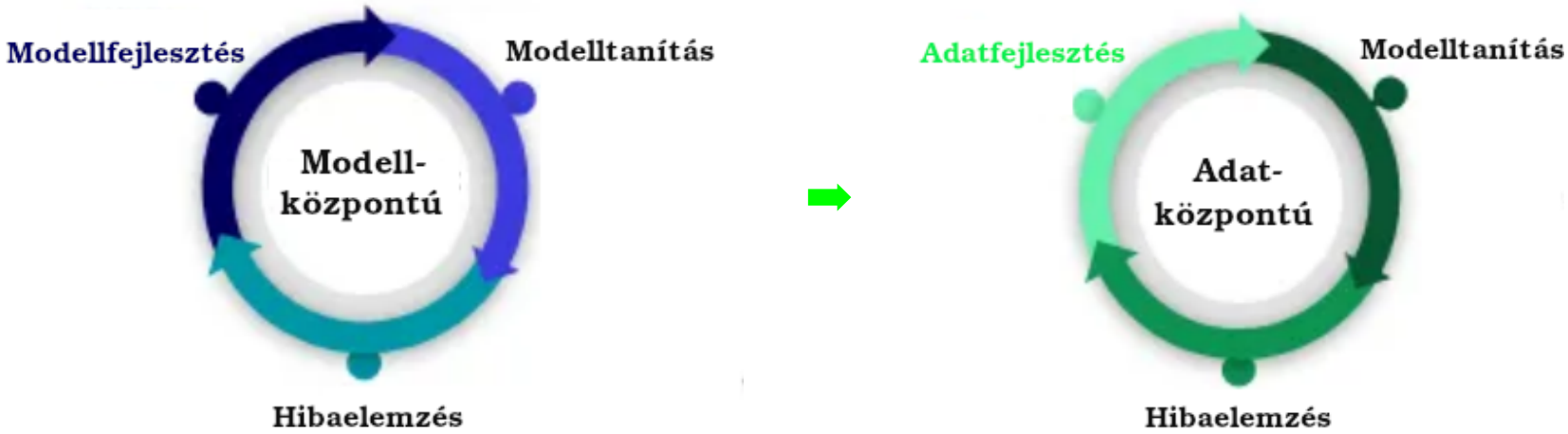
[a] Sambasivan et al. [2021]

Trendek



[3] https://www.zdnet.com/article/
the-state-of-ai-in-2021-machine-learning-in-production-mlops-and-data-centric-ai/

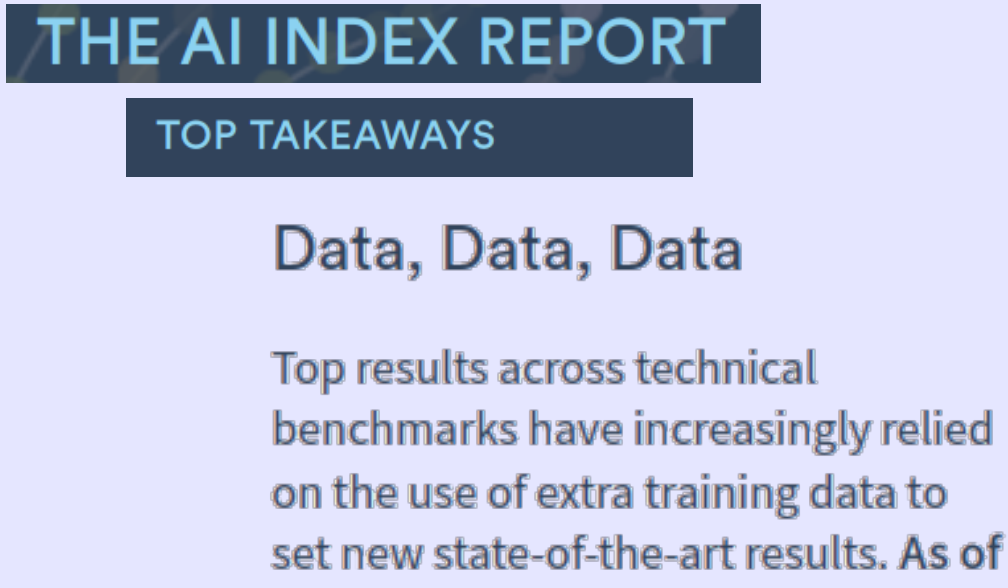
Trendek



[3] https://www.zdnet.com/article/
the-state-of-ai-in-2021-machine-learning-in-production-mlops-and-data-centric-ai/

Trendek

MI 2022–ben[4]



[4] https://aiindex.stanford.edu/report/

Trendek

big data ➡good data

adat, adaa, adat,
adatt, adt, adat,
data, adat, adta,
adat, tada

➡

adat, adat, adat

Adatközpontú MI

Mi az?

Az adatközpontú mesterséges intelligencia a MI-modellek építéséhez használt adatok szisztematikus fejlesztésének tudománya.

Súlyponteltolódás

- fejlett és széles körben elérhető MI-modellek $\Rightarrow$ modell változtatása csak minimális minőségjavulással jár
- valós felhasználás tanulsága $\Rightarrow$ az adat a fontosabb
- „Hogy építsük meg a legjobb modellt?” $\Rightarrow$ „Mivel tanítsuk a modellt?”

Elvárások

Milyen a jó adat?

1. jó minőségű
   - konzisztens
   - nem zajos
   - címkézési hibáktól mentes
2. reprezentatív
   - tükrözi a valós felhasználás körülményeit
   - nem tartalmaz más/felesleges adatot

Tartalom

Gépi fordítások

Forrás

Following their adoption of the euro, the main policy challenge for both Cyprus and Malta is to ensure the conduct of appropriate national economic policies in order to secure a high degree of sustainable convergence.

Gépi fordítások

Forrás

Following their adoption of the euro, the main policy challenge for both Cyprus and Malta is to ensure the conduct of appropriate national economic policies in order to secure a high degree of sustainable convergence.

α rendszer

Az euró bevezetése után Ciprus és Málta számára is a legfőbb gazdaságpolitikai kihívás az, hogy a fenntartható konvergencia magas fokának biztosítására megfelelő nemzeti gazdaságpolitikát folytassanak.

Gépi fordítások

Forrás

Following their adoption of the euro, the main policy challenge for both Cyprus and Malta is to ensure the conduct of appropriate national economic policies in order to secure a high degree of sustainable convergence.

α rendszer

Az euró bevezetése után Ciprus és Málta számára is a legfőbb gazdaságpolitikai kihívás az, hogy a fenntartható konvergencia magas fokának biztosítására megfelelő nemzeti gazdaságpolitikát folytassanak.

β rendszer

Az euró elfogadását követően Ciprus és Málta számára egyaránt a legfontosabb politikai kihívás, hogy a fenntartható fenntarthatóság magas szintjének biztosítása érdekében biztosítsa a megfelelő nemzeti gazdasági politikák végrehajtását.

Gépi fordítások

Forrás

Firazyr is given as a slow injection under the skin, preferably in the abdomen (tummy).

Gépi fordítások

Forrás

Firazyr is given as a slow injection under the skin, preferably in the abdomen (tummy).

α rendszer

Firaz Nifát lassú, lehetőleg a uralomban (tummy) adott injektáció formájában kapja.

Gépi fordítások

Forrás

Firazyr is given as a slow injection under the skin, preferably in the abdomen (tummy).

α rendszer

Firaz Nifát lassú, lehetőleg a uralomban (tummy) adott injektáció formájában kapja.

β rendszer

A Firazyr-t bőr alá fecskendezett lassú injekcióban kell beadni, lehetőleg a hasfalba szúrva.

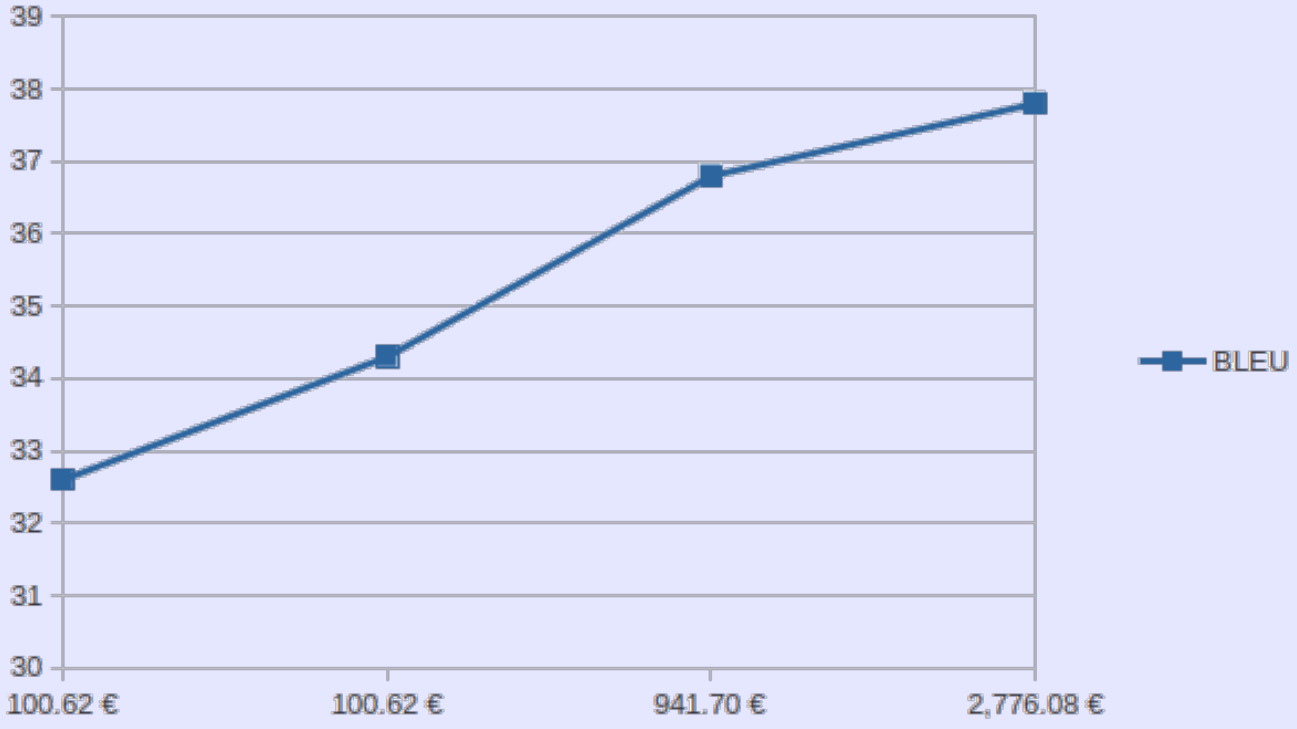
Tartalom

Modellközpontú módszerek

- komplex modell, mélyebb neurális háló
- komplex tanítási folyamat

Modellközpontúság

A komplexitás ára



Adatközpontú módszerek

1. tisztább adat, jobb fordítási modell (szűrés és javítás)
2. több adat, jobb fordítási modell (adatbővítés)
3. valós felhasználáshoz illeszkedő adat, jobb fordítási modell (szelektálás és finomhangolás)

Tisztább adat

Párhuzamos szegmensjelöltek

EN	HU
! "Yo no tengo Halo"?	*Hogy megvan-e?*
... after the slaying of onetime mobster...	*What's with your generation?*
... and on that note I'm done for now.	*De.... és amiért ez a bejegyzés most elkészült.*
... and sacrificed more than may humanly be expected.	*...több alázattal és áldozatkészséggel fogadta, mint az egy embertől egyáltalán várható.*
and that animal, Kuragin all he saw in her was a piece of sweet flesh to feed on.	*Tudom.*
and then slept for a few hours and watched TV. Just a beautiful job!	*Nagyszerű meló volt!*
64 blue, 19 red, 21 brown, 38 green.	*64 kék, 19 piros, 21 barna, 38 zöld.*
(9) Glyphosate is used as ester or salt.	*(9) A glifozát észterként vagy sóként használt.*
Government members shall swear an oath before Parliament.	*A Kormány tagja az Országgyűlés előtt esküt tesz.*
Imports of Atlantic swordfish originating in Sierra Leone are currently prohibited by Regulation (EC) No 828/2004.	*A Sierra Leonéból származó atlanti kardhal importját jelenleg tiltja a 828/ 2004/EK rendelet*

Tisztább adat

Párhuzamos szegmensjelöltek

nyelv

EN	HU
! "Yo no tengo Halo"?	*Hogy megvan-e?*
... after the slaying of onetime mobster...	*What's with your generation?*
... and on that note I'm done for now.	*De.... és amiért ez a bejegyzés most elkészült.*
... and sacrificed more than may humanly be expected.	*...több alázattal és áldozatkészséggel fogadta, mint az egy embertől egyáltalán várható.*
and that animal, Kuragin all he saw in her was a piece of sweet flesh to feed on.	*Tudom.*
and then slept for a few hours and watched TV. Just a beautiful job!	*Nagyszerű meló volt!*
64 blue, 19 red, 21 brown, 38 green.	*64 kék, 19 piros, 21 barna, 38 zöld.*
(9) Glyphosate is used as ester or salt.	*(9) A glifozát észterként vagy sóként használt.*
Government members shall swear an oath before Parliament.	*A Kormány tagja az Országgyűlés előtt esküt tesz.*
Imports of Atlantic swordfish originating in Sierra Leone are currently prohibited by Regulation (EC) No 828/2004.	*A Sierra Leonéból származó atlanti kardhal importját jelenleg tiltja a 828/ 2004/EK rendelet*

Tisztább adat

Párhuzamos szegmensjelöltek

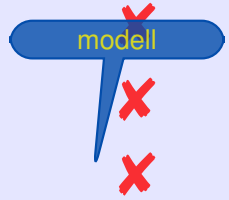
EN		HU
! "Yo no tengo Halo"?	✗	*Hogy megvan-e?*
... after the slaying of onetime mobster...	✗	*What's with your generation?*
... and on that note I'm done for now.		*De.... és amiért ez a bejegyzés most elkészült.*
... and sacrificed more than may humanly be expected.		*...több alázattal és áldozatkészséggel fogadta, mint az egy embertől egyáltalán várható.*
and that animal, Kuragin all he saw in her was a piece of sweet flesh to feed on.		*Tudom.*
and then slept for a few hours and watched TV. Just a beautiful job!		*Nagyszerű meló volt!*
64 blue, 19 red, 21 brown, 38 green.		*64 kék, 19 piros, 21 barna, 38 zöld.*
(9) Glyphosate is used as ester or salt.		*(9) A glifozát észterként vagy sóként használt.*
Government members shall swear an oath before Parliament.		*A Kormány tagja az Országgyűlés előtt esküt tesz.*
Imports of Atlantic swordfish originating in Sierra Leone are currently prohibited by Regulation (EC) No 828/2004.		*A Sierra Leonéból származó atlanti kardhal importját jelenleg tiltja a 828/ 2004/EK rendelet*

nyelv

Tisztább adat

Párhuzamos szegmensjelöltek

EN	HU
! "Yo no tengo Halo"?	*Hogy megvan-e?*
... after the slaying of onetime mobster...	*What's with your generation?*
... and on that note I'm done for now.	*De.... és amiért ez a bejegyzés most elkészült.*
... and sacrificed more than may humanly be expected.	*...több alázattal és áldozatkészséggel fogadta, mint az egy embertől egyáltalán várható.*
and that animal, Kuragin all he saw in her was a piece of sweet flesh to feed on.	*Tudom.*
and then slept for a few hours and watched TV. Just a beautiful job!	*Nagyszerű meló volt!*
64 blue, 19 red, 21 brown, 38 green.	*64 kék, 19 piros, 21 barna, 38 zöld.*
(9) Glyphosate is used as ester or salt.	*(9) A glifozát észterként vagy sóként használt.*
Government members shall swear an oath before Parliament.	*A Kormány tagja az Országgyűlés előtt esküt tesz.*
Imports of Atlantic swordfish originating in Sierra Leone are currently prohibited by Regulation (EC) No 828/2004.	*A Sierra Leonéból származó atlanti kardhal importját jelenleg tiltja a 828/ 2004/EK rendelet*



Tisztább adat

Párhuzamos szegmensjelöltek

EN		HU
! "Yo no tengo Halo"?	✗	*Hogy megvan-e?*
... after the slaying of onetime mobster...	✗	*What's with your generation?*
... and on that note I'm done for now.	✗	*De.... és amiért ez a bejegyzés most elkészült.*
... and sacrificed more than may humanly be expected.	✗	*...több alázattal és áldozatkészséggel fogadta, mint az egy embertől egyáltalán várható.*
and that animal, Kuragin all he saw in her was a piece of sweet flesh to feed on.		*Tudom.*
and then slept for a few hours and watched TV. Just a beautiful job!		*Nagyszerű meló volt!*
64 blue, 19 red, 21 brown, 38 green.		*64 kék, 19 piros, 21 barna, 38 zöld.*
(9) Glyphosate is used as ester or salt.		*(9) A glifozát észterként vagy sóként használt.*
Government members shall swear an oath before Parliament.		*A Kormány tagja az Országgyűlés előtt esküt tesz.*
Imports of Atlantic swordfish originating in Sierra Leone are currently prohibited by Regulation (EC) No 828/2004.		*A Sierra Leonéból származó atlanti kardhal importját jelenleg tiltja a 828/ 2004/EK rendelet*

modell

Tisztább adat

Párhuzamos szegmensjelöltek

EN		HU
! "Yo no tengo Halo"?	✗	*Hogy megvan-e?*
... after the slaying of onetime mobster...	✗	*What's with your generation?*
... and on that note I'm done for now.	✗	*De.... és amiért ez a bejegyzés most elkészült.*
... and sacrificed more than may humanly be expected.	✗	*...több alázattal és áldozatkészséggel fogadta, mint az egy embertől egyáltalán várható.*
and that animal, Kuragin all he saw in her was a piece of sweet flesh to feed on.	✗	*Tudom.*
and then slept for a few hours and watched TV. Just a beautiful job!		*Nagyszerű meló volt!*
64 blue, 19 red, 21 brown, 38 green.		*64 kék, 19 piros, 21 barna, 38 zöld.*
(9) Glyphosate is used as ester or salt.		*(9) A glifozát észterként vagy sóként használt.*
Government members shall swear an oath before Parliament.		*A Kormány tagja az Országgyűlés előtt esküt tesz.*
Imports of Atlantic swordfish originating in Sierra Leone are currently prohibited by Regulation (EC) No 828/2004.		*A Sierra Leonéból származó atlanti kardhal importját jelenleg tiltja a 828/ 2004/EK rendelet*

hossz

Tisztább adat

Párhuzamos szegmensjelöltek

EN		HU
! "Yo no tengo Halo"?	✗	*Hogy megvan-e?*
... after the slaying of onetime mobster...	✗	*What's with your generation?*
... and on that note I'm done for now.	✗	*De.... és amiért ez a bejegyzés most elkészült.*
... and sacrificed more than may humanly be expected.	✗	*...több alázattal és áldozatkészséggel fogadta, mint az egy embertől egyáltalán várható.*
and that animal, Kuragin all he saw in her was a piece of sweet flesh to feed on.	✗	*Tudom.*
and then slept for a few hours and watched TV. Just a beautiful job!	✗	*Nagyszerű meló volt!*
64 blue, 19 red, 21 brown, 38 green.		*64 kék, 19 piros, 21 barna, 38 zöld.*
(9) Glyphosate is used as ester or salt.		*(9) A glifozát észterként vagy sóként használt.*
Government members shall swear an oath before Parliament.		*A Kormány tagja az Országgyűlés előtt esküt tesz.*
Imports of Atlantic swordfish originating in Sierra Leone are currently prohibited by Regulation (EC) No 828/2004.		*A Sierra Leonéból származó atlanti kardhal importját jelenleg tiltja a 828/ 2004/EK rendelet*

hossz

Tisztább adat

Párhuzamos szegmensjelöltek

EN		HU
! "Yo no tengo Halo"?	✗	*Hogy megvan-e?*
... after the slaying of onetime mobster...	✗	*What's with your generation?*
... and on that note I'm done for now.	✗	*De.... és amiért ez a bejegyzés most elkészült.*
... and sacrificed more than may humanly be expected.	✗	*...több alázattal és áldozatkészséggel fogadta, mint az egy embertől egyáltalán várható.*
and that animal, Kuragin all he saw in her was a piece of sweet flesh to feed on.	✗	*Tudom.*
and then slept for a few hours and watched TV. Just a beautiful job!	✗	*Nagyszerű meló volt!*
64 blue, 19 red, 21 brown, 38 green.	✗	*64 kék, 19 piros, 21 barna, 38 zöld.*
(9) Glyphosate is used as ester or salt.		*(9) A glifozát észterként vagy sóként használt.*
Government members shall swear an oath before Parliament.		*A Kormány tagja az Országgyűlés előtt esküt tesz.*
Imports of Atlantic swordfish originating in Sierra Leone are currently prohibited by Regulation (EC) No 828/2004.		*A Sierra Leonéból származó atlanti kardhal importját jelenleg tiltja a 828/ 2004/EK rendelet*

betű-szám arány

Tisztább adat

Párhuzamos szegmensjelöltek

EN		HU
! "Yo no tengo Halo"?	✗	*Hogy megvan-e?*
... after the slaying of onetime mobster...	✗	*What's with your generation?*
... and on that note I'm done for now.	✗	*De.... és amiért ez a bejegyzés most elkészült.*
... and sacrificed more than may humanly be expected.	✗	*...több alázattal és áldozatkészséggel fogadta, mint az egy embertől egyáltalán várható.*
and that animal, Kuragin all he saw in her was a piece of sweet flesh to feed on.	✗	*Tudom.*
and then slept for a few hours and watched TV. Just a beautiful job!	✗	*Nagyszerű meló volt!*
64 blue, 19 red, 21 brown, 38 green.	✗	*64 kék, 19 piros, 21 barna, 38 zöld.*
(9) Glyphosate is used as ester or salt.	✔	*(9) A glifozát észterként vagy sóként használt.*
Government members shall swear an oath before Parliament.		*A Kormány tagja az Országgyűlés előtt esküt tesz.*
Imports of Atlantic swordfish originating in Sierra Leone are currently prohibited by Regulation (EC) No 828/2004.		*A Sierra Leonéból származó atlanti kardhal importját jelenleg tiltja a 828/ 2004/EK rendelet*

modell

Tisztább adat

Párhuzamos szegmensjelöltek

EN		HU
! "Yo no tengo Halo"?	✗	*Hogy megvan-e?*
... after the slaying of onetime mobster...	✗	*What's with your generation?*
... and on that note I'm done for now.	✗	*De.... és amiért ez a bejegyzés most elkészült.*
... and sacrificed more than may humanly be expected.	✗	*...több alázattal és áldozatkészséggel fogadta, mint az egy embertől egyáltalán várható.*
and that animal, Kuragin all he saw in her was a piece of sweet flesh to feed on.	✗	*Tudom.*
and then slept for a few hours and watched TV. Just a beautiful job!	✗	*Nagyszerű meló volt!*
64 blue, 19 red, 21 brown, 38 green.	✗	*64 kék, 19 piros, 21 barna, 38 zöld.*
(9) Glyphosate is used as ester or salt.	✓	*(9) A glifozát észterként vagy sóként használt.*
Government members shall swear an oath before Parliament.	✓	*A Kormány tagja az Országgyűlés előtt esküt tesz.*
Imports of Atlantic swordfish originating in Sierra Leone are currently prohibited by Regulation (EC) No 828/2004.		*A Sierra Leonéból származó atlanti kardhal importját jelenleg tiltja a 828/ 2004/EK rendelet*

modell

Tisztább adat

Párhuzamos szegmensjelöltek

EN		HU
! "Yo no tengo Halo"?	✗	*Hogy megvan-e?*
... after the slaying of onetime mobster...	✗	*What's with your generation?*
... and on that note I'm done for now.	✗	*De.... és amiért ez a bejegyzés most elkészült.*
... and sacrificed more than may humanly be expected.	✗	*...több alázattal és áldozatkészséggel fogadta, mint az egy embertől egyáltalán várható.*
and that animal, Kuragin all he saw in her was a piece of sweet flesh to feed on.	✗	*Tudom.*
and then slept for a few hours and watched TV. Just a beautiful job!	✗	*Nagyszerű meló volt!*
64 blue, 19 red, 21 brown, 38 green.	✗	*64 kék, 19 piros, 21 barna, 38 zöld.*
(9) Glyphosate is used as ester or salt.	✓ modell	*(9) A glifozát észterként vagy sóként használt.*
Government members shall swear an oath before Parliament.	✓	*A Kormány tagja az Országgyűlés előtt esküt tesz.*
Imports of Atlantic swordfish originating in Sierra Leone are currently prohibited by Regulation (EC) No 828/2004.	✓	*A Sierra Leonéból származó atlanti kardhal importját jelenleg tiltja a 828/ 2004/EK rendelet*

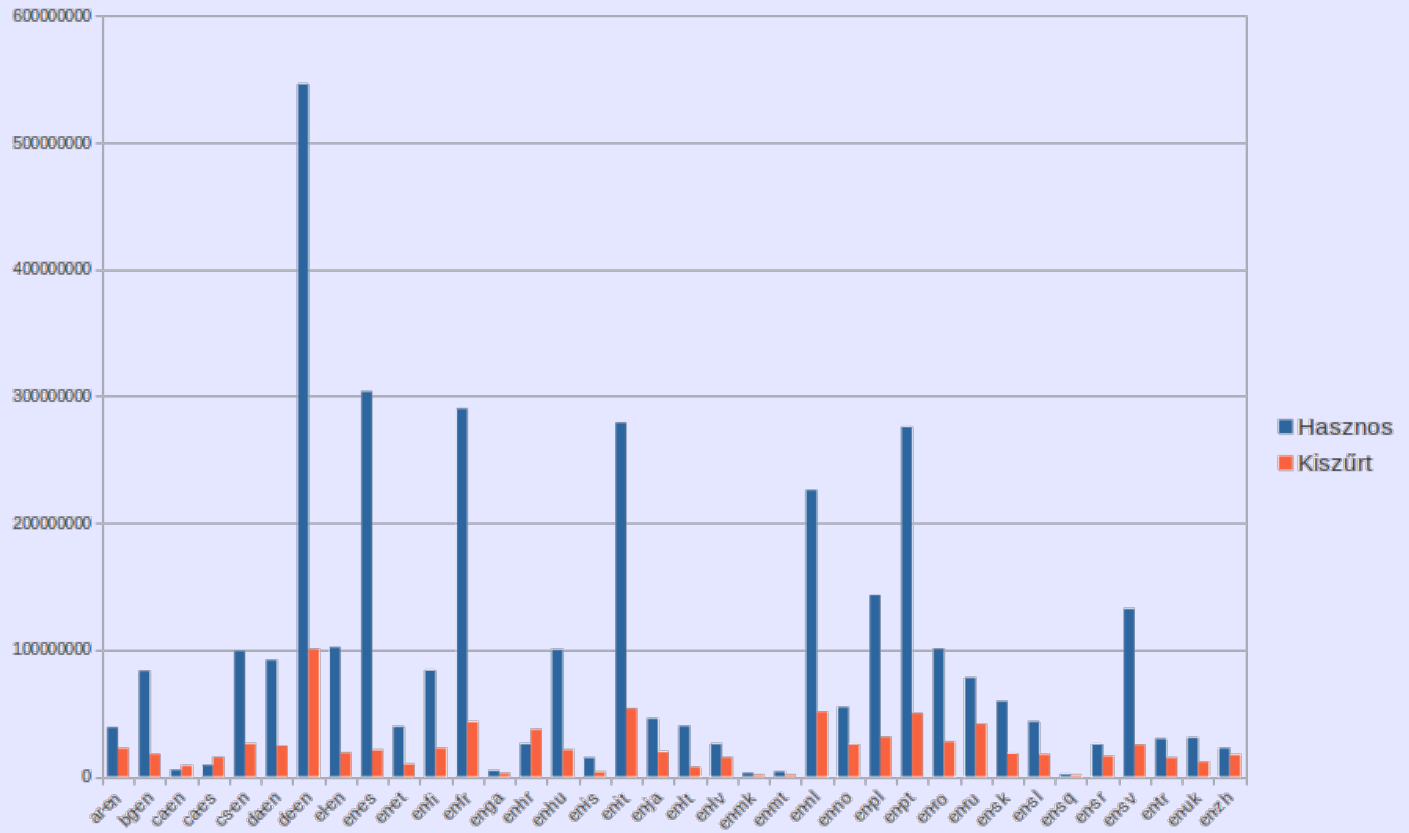
Tisztább adat

Párhuzamos szegmensjelöltek

EN		HU
! "Yo no tengo Halo"?	✗	*Hogy megvan-e?*
... after the slaying of onetime mobster...	✗	*What's with your generation?*
... and on that note I'm done for now.	✗	*De.... és amiért ez a bejegyzés most elkészült.*
... and sacrificed more than may humanly be expected.	✗	*...több alázattal és áldozatkészséggel fogadta, mint az egy embertől egyáltalán várható.*
and that animal, Kuragin all he saw in her was a piece of sweet flesh to feed on.	✗	*Tudom.*
and then slept for a few hours and watched TV. Just a beautiful job!	✗	*Nagyszerű meló volt!*
64 blue, 19 red, 21 brown, 38 green.	✗	*64 kék, 19 piros, 21 barna, 38 zöld.*
(9) Glyphosate is used as ester or salt.	✓	*(9) A glifozát észterként vagy sóként használt.*
Government members shall swear an oath before Parliament.	✓	*A Kormány tagja az Országgyűlés előtt esküt tesz.*
Imports of Atlantic swordfish originating in Sierra Leone are currently prohibited by Regulation (EC) No 828/2004.	✓	*A Sierra Leonéból származó atlanti kardhal importját jelenleg tiltja a 828/ 2004/EK rendelet*

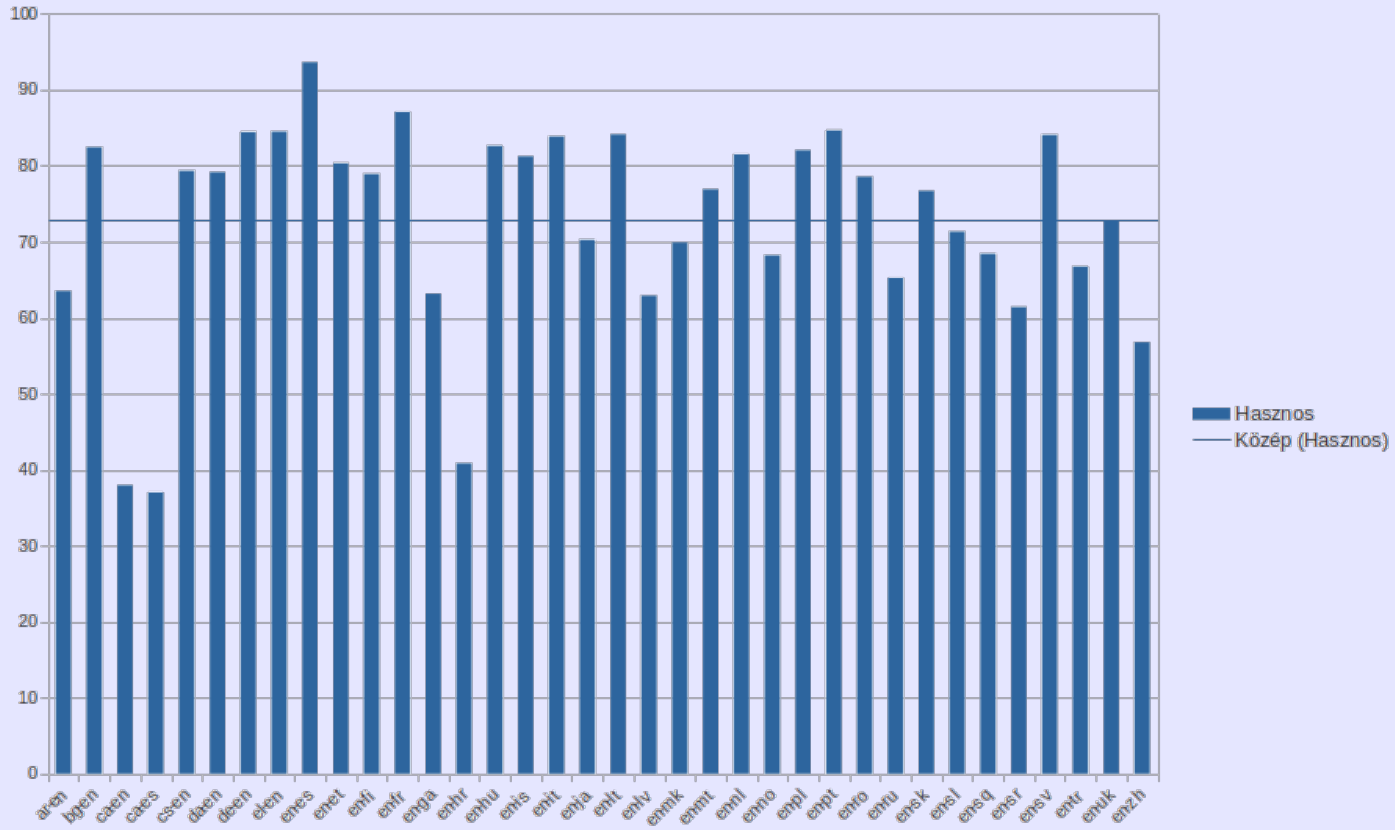
Tisztább adataink

Hasznos és haszontalan szegmensek nyelvpárok szerint



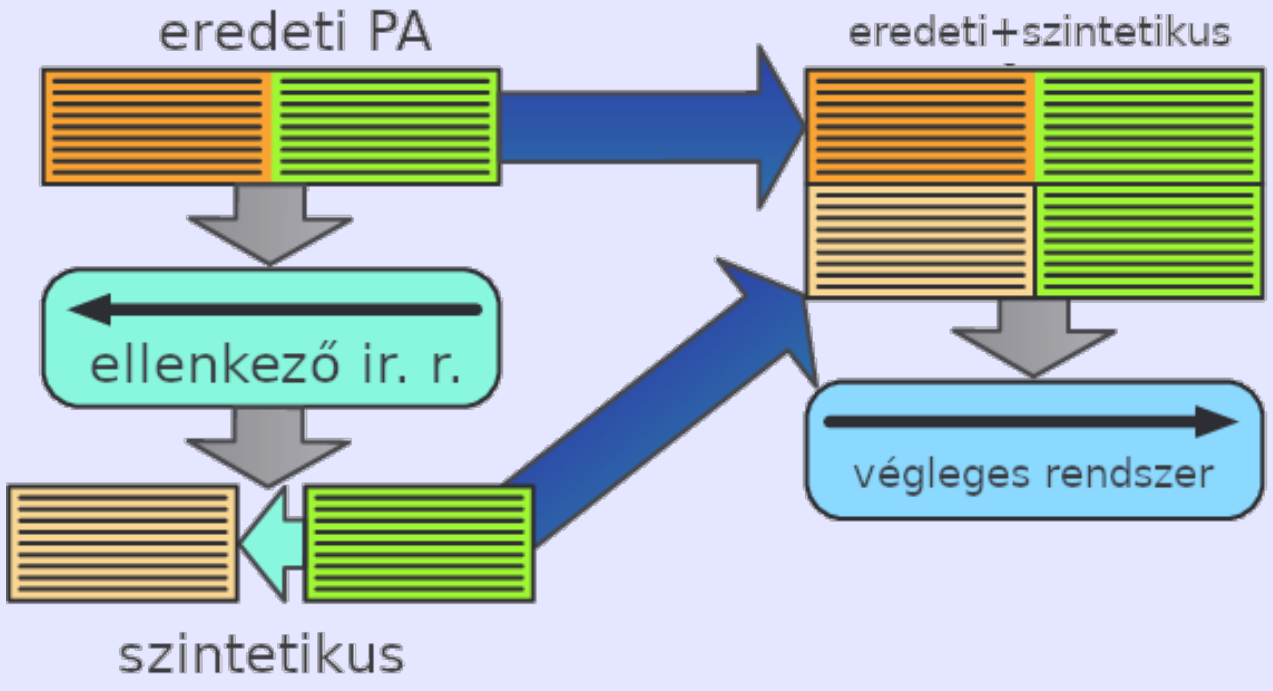
Tisztább adataink

Hasznos szegmensek aránya nyelvpárok szerint



Több adat: adatbővítés

Visszafordítás (back-translation)[5]



[5] Forrás: Hoang et al. [2018]

Visszafordítás (back-translation)

Példák

EN	←	HU
	?	*Az adózási szempontból nem együttműködő országok és területek európai uniós jegyzékének legutóbbi időszakos felülvizsgálata során az EU felvette Dominikát a jegyzékbe, Barbadost pedig törölte a jegyzékből.*
	?	*A dokumentációnak tartalmaznia kell a készterméken a felszabadításkor végzett ellenőrző vizsgálatokkal és validálásukkal kapcsolatos adatokat.*
	?	*Kitűnő példa erre a lakóövezetekben, valamint a sűrű kerékpáros és gyalogos forgalommal jellemzett övezetekben a 30 km/órás legnagyobb sebesség előírása.*
	?	*Az utóbbi három vélemény csak angol nyelven kerül megvitatásra.*

Visszafordítás (back-translation)

Példák

EN	HU
In its periodical review of the EU list of non-cooperative tax jurisdictions, the EU decided to include Dominica and remove Barbados from the list.	*Az adózási szempontból nem együttműködő országok és területek európai uniós jegyzékének legutóbbi időszakos felülvizsgálata során az EU felvette Dominikát a jegyzékbe, Barbadost pedig törölte a jegyzékből.*
The dossier shall include particulars relating to control tests on the finished product at release and their validation.	*A dokumentációnak tartalmaznia kell a készterméken a felszabadításkor végzett ellenőrző vizsgálatokkal és validálásukkal kapcsolatos adatokat.*
A very good example is the maximum speeds of 30 km/h in residential areas and areas where there are high levels of cyclists and pedestrians.	*Kitűnő példa erre a lakóövezetekben, valamint a sűrű kerékpáros és gyalogos forgalommal jellemzett övezetekben a 30 km/órás legnagyobb sebesség előírása.*
The last three opinions are to be dealt with in English language only.	*Az utóbbi három vélemény csak angol nyelven kerül megvitatásra.*

Visszafordítás (back-translation)

Példák

EN	←	HU
! "Yo no tengo Halo"?	✘	*Hogy megvan-e?*
... and on that note I'm done for now.	✘	*De.... és amiért ez a bejegyzés most elkészült.*
... and sacrificed more than may humanly be expected.	✘	*...több alázattal és áldozatkészséggel fogadta, mint az egy embertől egyáltalán várható.*
and that animal, Kuragin all he saw in her was a piece of sweet flesh to feed on.	✘	*Tudom.*
and then slept for a few hours and watched TV. Just a beautiful job!	✘	*Nagyszerű meló volt!*

Visszafordítás (back-translation)

Példák

EN		HU
Do you have it?	←	*Hogy megvan-e?*
But... and that's why this post is now completed.	←	*De.... és amiért ez a bejegyzés most elkészült.*
he received it with more humility and sacrifice than one man could ever expect.	←	*...több alázattal és áldozatkészséggel fogadta, mint az egy embertől egyáltalán várható.*
I know it.	←	*Tudom.*
That was a great job.	←	*Nagyszerű meló volt!*

Gépi fordítás

Forrás

The staff committee chairs should receive, at the same time as Directors General, the Bureau's agendas and summaries of its decisions.

Gépi fordítás

Forrás

The staff committee chairs should receive, at the same time as Directors General, the Bureau's agendas and summaries of its decisions.

Alaprendszer

A személyzeti bizottságok elnökei a főigazgatókkal egy időben megkapják az Elnökség napirendjét és határozatainak összefoglalóit.

Gépi fordítás

Forrás

Finally, the janitor took over the chairs from the delivery service.

Alaprendszer

Gépi fordítás

Forrás

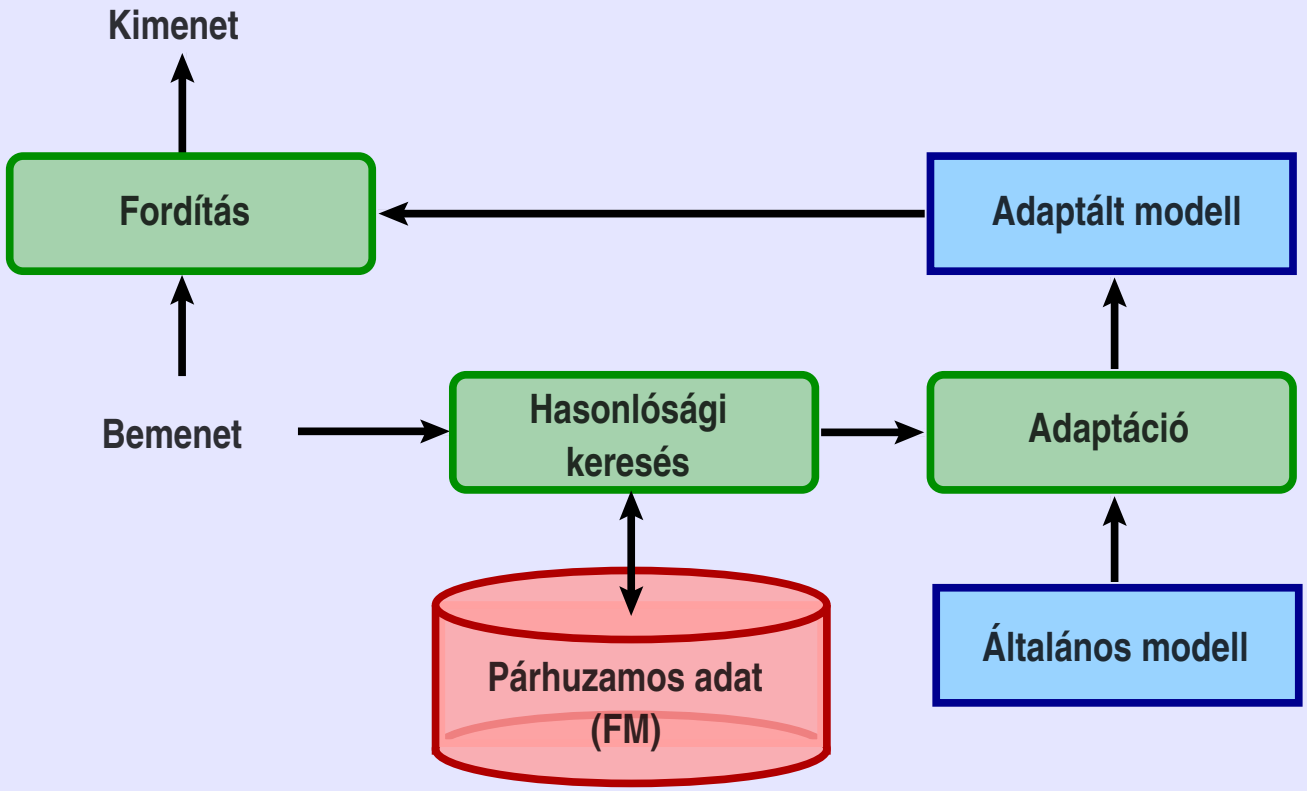
Finally, the janitor took over the chairs from the delivery service.

Alaprendszer

Végül az iroda átveszi az elnököket a kézbesítési szolgáltatástól.

Fordítási feladathoz illeszkedő adat

Bemeneti adat által vezérelt modelladaptáció



Gépi fordítás

Forrás

Finally, the janitor took over the chairs from the delivery service.

Alaprendszer

Végül az iroda átveszi az elnököket a kézbesítési szolgáltatástól.

Adaptált rendszer

Gépi fordítás

Forrás

Finally, the janitor took over the chairs from the delivery service.

Alaprendszer

Végül az iroda átveszi az elnököket a kézbesítési szolgáltatástól.

Adaptált rendszer

Végül a gondnok átvette a székeket a kézbesítő szolgálattól.

Tartalom

Gépi elő- és utófeldolgozás

Normalizálás

- írásjelek, tizedespont és ezreselválasztó, gondolatjel, listák

Gépi elő- és utófeldolgozás

Truecasing

- COMMISSION STATEMENT...
- Comission Statement. . .
- commission statement...

Gépi elő- és utófeldolgozás

Tokenizálás

- 2. cikkének (1) bekezdése előírja, hogy az Európai Központi Bank („EKB”)
- 2 _. cikkének (_ 1 _) bekezdése előírja _, hogy az Európai Központi Bank (_ _„_ EKB _”_ _)

Gépi elő- és utófeldolgozás

Helykitöltők/placeholderek

- Commission Regulation (EC) No 1517/95 of 29 June 1995
- 1995. június 29-i 1517/95/EK bizottsági rendelet
- Commission Regulation (EC) No _num1/_num2 of _num3 June _num4
- _num4. június _num3-i _num1 /_num2 /EK bizottsági rendelet

Szótárépítés

- Valós szókincs (tokenizálás/truecasing/helykitöltők után) kezelhetetlenül nagy
- Praktikus jellemző (kezelhető) szótárméret: 32000 (forrás és célnyelv összesen)
- (Mohó) szegmentálás a leggyakrabban előforduló szósorozatokra

Szótárépítés

- Valós szókincs (tokenizálás/truecasing/helykitöltők után) kezelhetetlenül nagy
- Praktikus jellemző (kezelhető) szótárméret: 32000 (forrás és célnyelv összesen)
- (Mohó) szegmentálás a leggyakrabban előforduló szósorozatokra

Példa

```
Az ESZSZ és az UNICE által a határozott ideig tartó
munkaviszonyról kötött keretmegállapodáson, mielőtt a
Bíróság kérdéseit tárgyalták volna
```


Szótárépítés

- Valós szókincs (tokenizálás/truecasing/helykitöltők után) kezelhetetlenül nagy
- Praktikus jellemző (kezelhető) szótárméret: 32000 (forrás és célnyelv összesen)
- (Mohó) szegmentálás a leggyakrabban előforduló szósorozatokra

Példa

```
Az ESZ@@ SZ és az UN@@ ICE által a határozott ideig
tartó munkaviszony@@ ról kötött keret@@ megállapodáson,
mielőtt a Bíróság kérdés@@ eit tárgyal@@ ták volna
```


Tartalom

Minőségbecslés – QE

A fordítás minőségének automatikus becslése refrenciafordítás nélkül.

- szószintű: a fordítás minden szavának felcímkézése (+/-)
- szegmensszintű: a szegmenshez a fordítás minőségét jelző mérőszám rendelése

Adaptív fordítás

- Fordítómotor mondatonkénti tanítása utószerkesztéskor
- Tanulóanyag: forrás- kijavított célszegmenspár
- Modelparaméterek beállítása (learning rate változtatása)
- + mondatonkénti javulás
- - folyamatos GPU használat
- tréning + fordítás idő $<$ 1 s/szegmens

Összefoglaló

- a neurális fordítórendszerek paradigmaváltást és minőségi ugrást hoztak a gépi fordításban (is)
- már egy egyszerű modell is jó teljesítményre képes (de a SOTA modellek már nem egyszerűek...)
- a nyers erő hasznos de sokba kerül
- az adatok megfelelő menedzselése olcsóbb és hatékonyabb

Házi feladat

1. Olvass utána:
   - nem felügyelt gépi fordítás (unsupervised MT)
   - katasztrofális felejtés (catastrophic forgetting)
2. Hogy építenénk egy olyan fordítási modellt, ami több nyelvről képes több nyelvre fordítani? Milyen nehézségek merülhetnek fel?
3. Hogy lehetne előállítani a dinamikusan adaptált modellhez szükséges adathalmazt?
4. Milyen módszerekkel szabadulhatunk meg a túlságosan zajos (nem) párhuzamos adatoktól?
5. Hogyan mérhetnénk a gépi fordítás minőségét?

Vége

Hivatkozások I

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2703. URL https://www.aclweb.org/anthology/W18-2703.

Attila Novák, Borbala Novák, and Gábor Prószéky. Segíthetnek-e a szóbeágyazási modellek a társadalomtudósoknak? *MAGYAR TUDOMÁNY*, 179:945–954, 2018. ISSN 0025-0325. doi: 10.1556/2065.179.2018.7.3. URL https://mersz.hu/mod/object.php?objazonosito=matud_f8922_i5.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen K. Paritosh, and Lora Aroyo. "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker, editors, *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pages 39:1–39:15. ACM, 2021. doi: 10.1145/3411764.3445518. URL https://doi.org/10.1145/3411764.3445518.