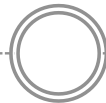




A gépi fordítás kiértékelése



Yang Zijian Győző

Témavezető:

Dr. Prószéky Gábor

Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar

Roska Tamás Műszaki és Természettudományi Doktori Iskola



Összefoglaló

- Bevezető
- Gépi fordítás
- Gépi fordítás kiértékelése
- Összegzés



Google



Bejelentkezés

Fordító

Azonnali fordítás kikapcsolása



angol magyar német Nyelvfelismerés ▾



magyar angol német ▾

Fordítás

I am a spendthrift at heart.



28/5000

Én egy pazarló a szíve.



Google



Bejelentkezés

Fordító

Azonnali fordítás kikapcsolása



angol magyar német Nyelvfelismerés ▾



magyar angol német ▾

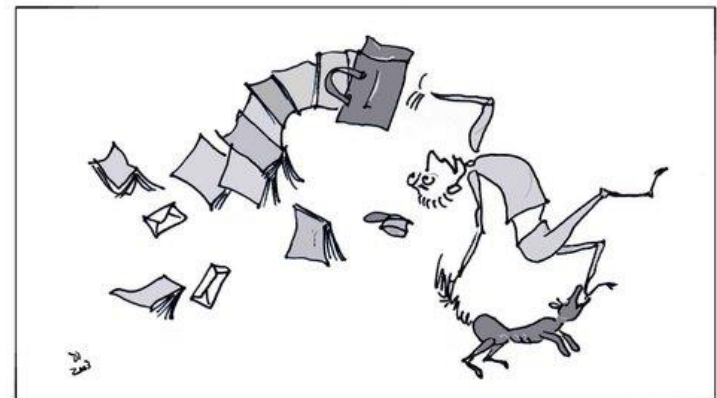
Fordítás

The dog was bitten by the postman. ×



34/5000

A kutya megharapta a postást.





- Cél: megértés, információkinyerés, nem a tökéletes fordítás



- Cél: megértés, információkinyerés, nem a tökéletes fordítás
- Főbb módszerek:
 - Példa alapú / fordítómemórián alapuló gépi fordítás

- Cél: megértés, információkinyerés, nem a tökéletes fordítás
- Főbb módszerek:
 - Példa alapú / fordítómemórián alapuló gépi fordítás



- Cél: megértés, információkinyerés, nem a tökéletes fordítás
- Főbb módszerek:
 - Példa alapú / fordítómemórián alapuló gépi fordítás



- Szabály alapú gépi fordítás
 - a felnőtt: szabályok megtanulása és szótár

- Cél: megértés, információkinyerés, nem a tökéletes fordítás
- Főbb módszerek:
 - Példa alapú / fordítómemórián alapuló gépi fordítás



- Szabály alapú gépi fordítás
 - a felnőtt: szabályok megtanulása és szótár





- Főbb módszerek:
 - Statisztikai gépi fordítás
 - a gyerek: példák megfigyelése
 - párhuzamos korpusz

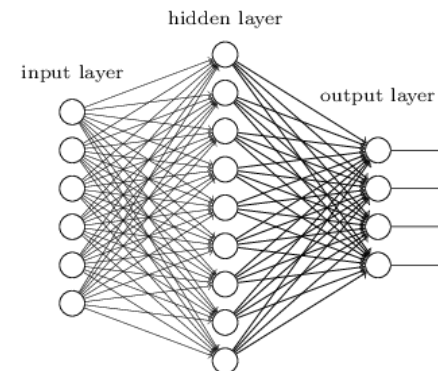
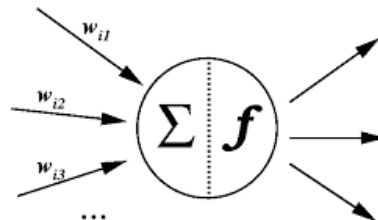
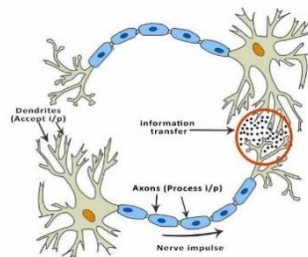
- Főbb módszerek:
 - Statisztikai gépi fordítás
 - a gyerek: példák megfigyelése
 - párhuzamos korpusz



- Főbb módszerek:
 - Statisztikai gépi fordítás
 - a gyerek: példák megfigyelése
 - párhuzamos korpusz



- Neurális gépi fordítás
 - mesterséges neurális hálózat



- Főbb módszerek:
 - Statisztikai gépi fordítás
 - a gyerek: példák megfigyelése
 - párhuzamos korpusz



- Neurális gépi fordítás
 - mesterséges neurális hálózat





- Gépi fordítás utómunkát végző szakemberek számára.
- Gépi fordító rendszerek kimenetének vegyítése.
- Gépi fordítást kiértékelő szakemberek számára.
- Kiszűrni a használhatatlan fordításokat.
- Figyelmeztetni a felhasználót a hibás szövegrészletekre.



- Emberi kiértékelés
 - Legpontosabb.
 - Drága, lassú.
- Gépi kiértékelés
 - Referenciafordítással történő kiértékelés
 - Drága, nem valós idejű kiértékelés.
 - Minőségbecslés (Quality Estimation)
 - Nem használ referenciafordítást.
 - Valós idejű kiértékelés.
 - Magas korreláció az emberi kiértékeléssel.



- Tartalomhűség (adequacy)
 - jelentésben mennyire jó
- Gördülékenység, nyelvhelyesség (fluency)
 - a célnyelvi mondat nyelvtanilag mennyire helyes
- Elfogadhatóság



- Pontosság
 - $\text{pontosság} = \frac{|\text{metszet}|}{|\text{eredmény}|}$
- Fedés
 - $\text{fedés} = \frac{|\text{metszet}|}{|\text{cél}|}$
- F-mérték
 - pontosság és fedés harmonikus közepe, átlaga
- Utószerkesztés (post-editing effort)
- Használhatóság
 - OK / BAD

- **Forrás:** Ez egy útmutató, amely biztosítja, hogy a hadsereg mindig betartsa a párt utasításait.
- **Gépi fordítás:** *It is a guide to action which ensures that the military always obeys the commands of the party.*
- **1. referencia:** *It is a guide to action that ensures that the military will forever heed Party commands.*
- **2. referencia:** *It is the guiding principle which guarantees the military forces always being under the command of the Party.*
- **3. referencia:** *It is the practical guide for the army always to heed the directions of the party.*



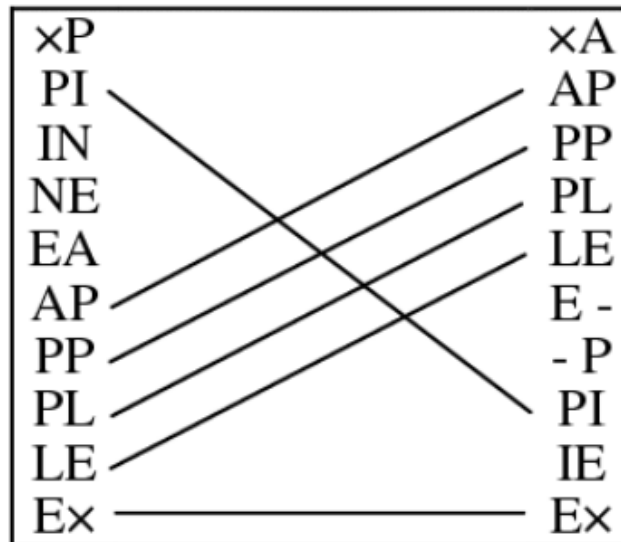
- Automatikus kiértékelő módszerek.
- Nincsen szükség emberi beavatkozásra.
- Nem kell ismerni a két nyelvet.
- Hasonlóságot és eltérést mér a fordítás és a referenciafordítás között.
- Módszerek:
 - BLEU, NIST, METEOR, LEPOR, ROGUE, RIBES
 - WER, TER, HTER

- BiLingual Evaluation Understudy
- pontosság: a lefordított mondatokban lévő szavak és kifejezések mennyire illeszkednek pontosan a referenciafordításhoz.
- **1. fordítás:** *It is a guide to action which ensures that the military always obeys the commands of the party*
- **2. fordítás:** *It is to insure the troops forever hearing the activity guidebook that party direct*
- **1. referencia:** *It is a guide to action that ensures that the military will forever heed Party commands*
- **2. referencia:** *It is the guiding principle which guarantees the military forces always being under the command of the Party*
- **3. referencia:** *It is the practical guide for the army always to heed the directions of the party*



- Tartalomhűség: unigram pontosság
- Gördülékenység: n-gram ($n > 1$) pontosság
- Az algoritmus az n-gramokból számolt értékek súlyozott átlagát (mértani átlag) adja eredményül
- Előnyök:
 - Gyors, olcsó, automatikus.
- Hátrányok:
 - Kevésbé érzékeny a szórendi átalakításra.
 - Különböző toldalék esetén eltérő BLEU érték.

- Karakterszintű n-gramok
- Ragozó nyelvek esetén előnyös
- F-mérték számolása
- Emberi utómunka számára nyújthat segítséget (post-edit)





- NIST
 - információtartalom
- ROGUE
 - Fedést számol
- METEOR
 - Szótövesítő, szinonimák kezelése
- LEPOR
 - Pozíció, F-mérték
- RIBES
 - Szórendhasonlóság korrelációval



- Translation Edit/Error Rate

$$TER = \frac{\text{javítások száma}}{\text{referenciafordítás szavainak átlagos száma}}$$

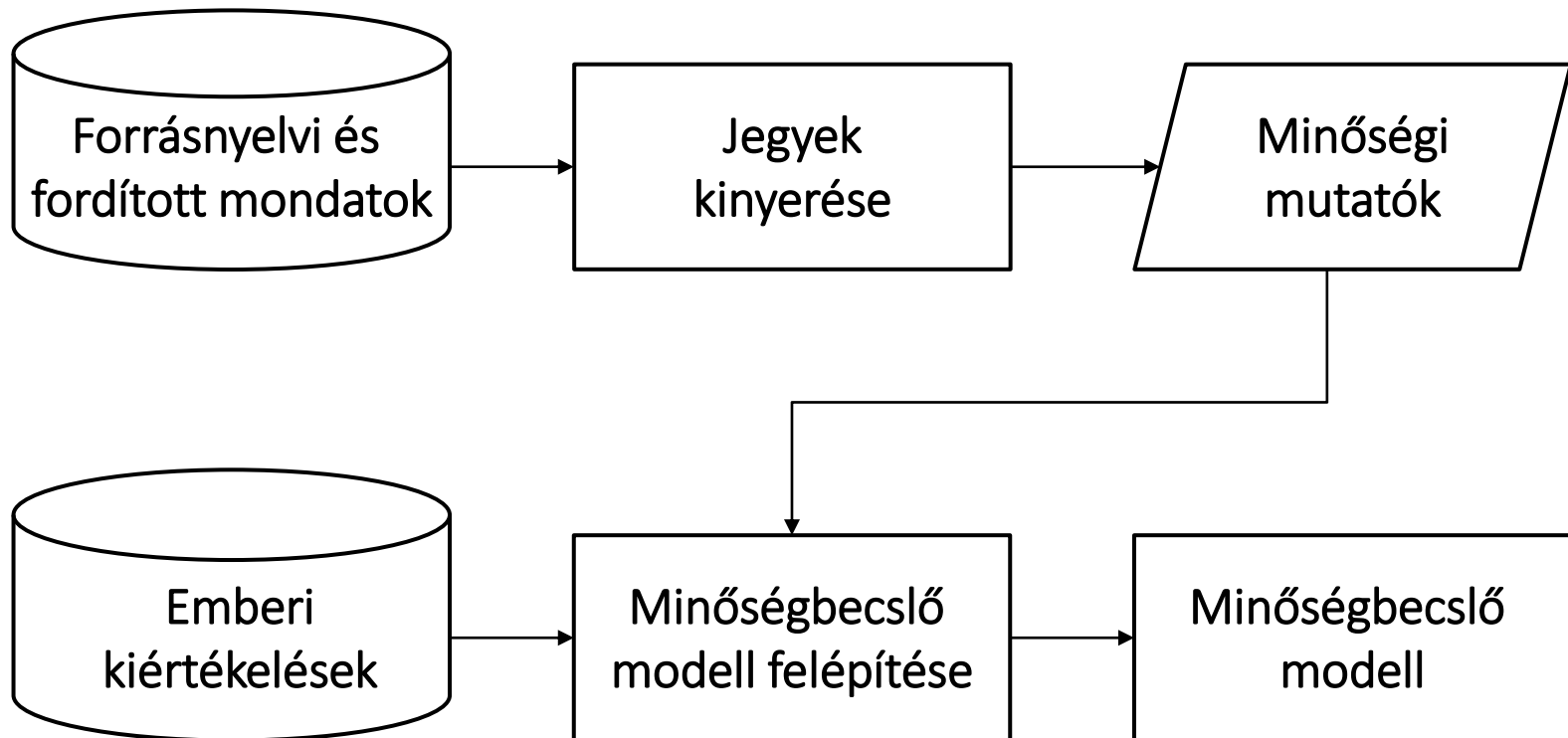
- Javítások:
 - beszúrás, törlés, csere, eltolás
- Nem kezeli a szemantikai problémákat.
- Kevesebb javítással létrehozható hasonló jelentésű mondat.

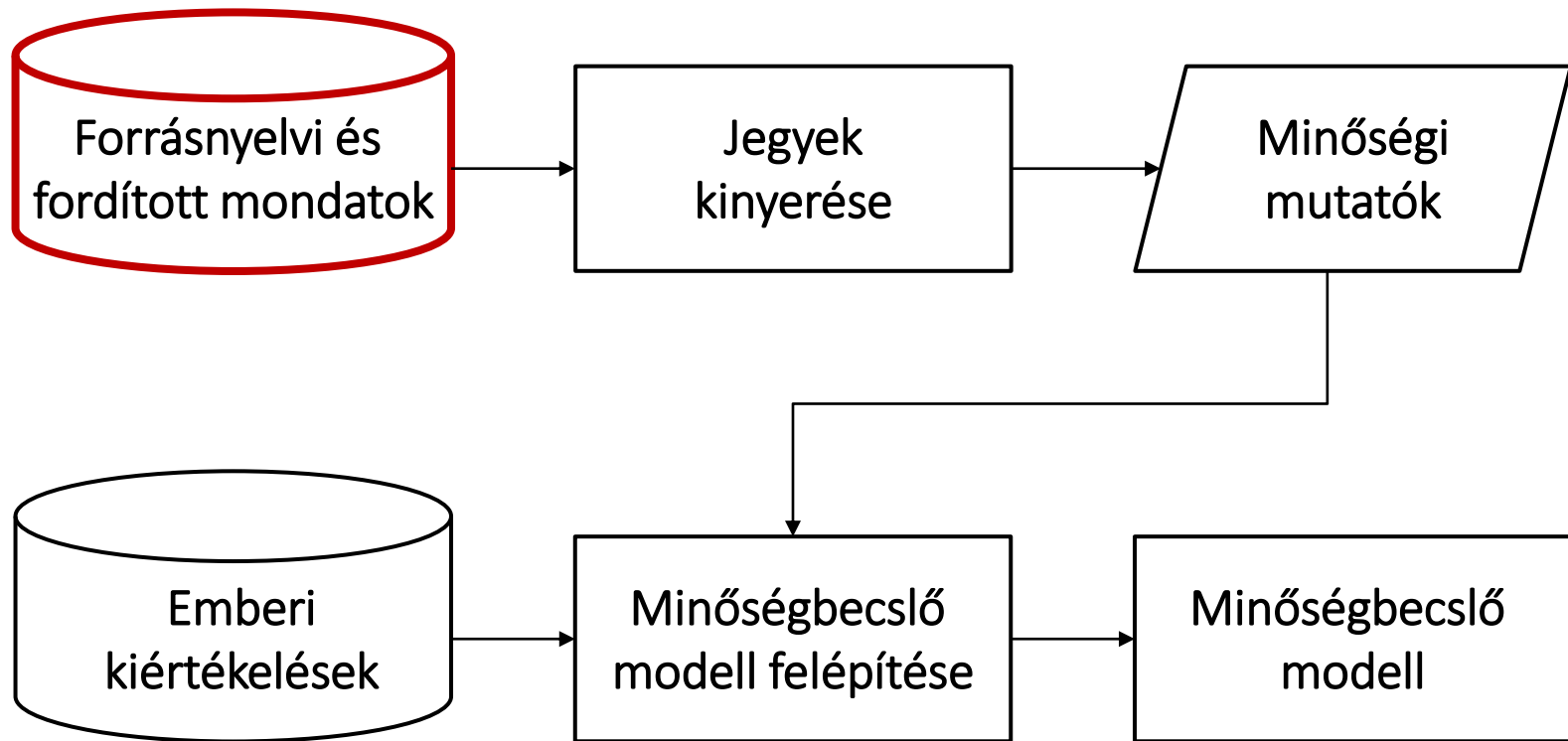


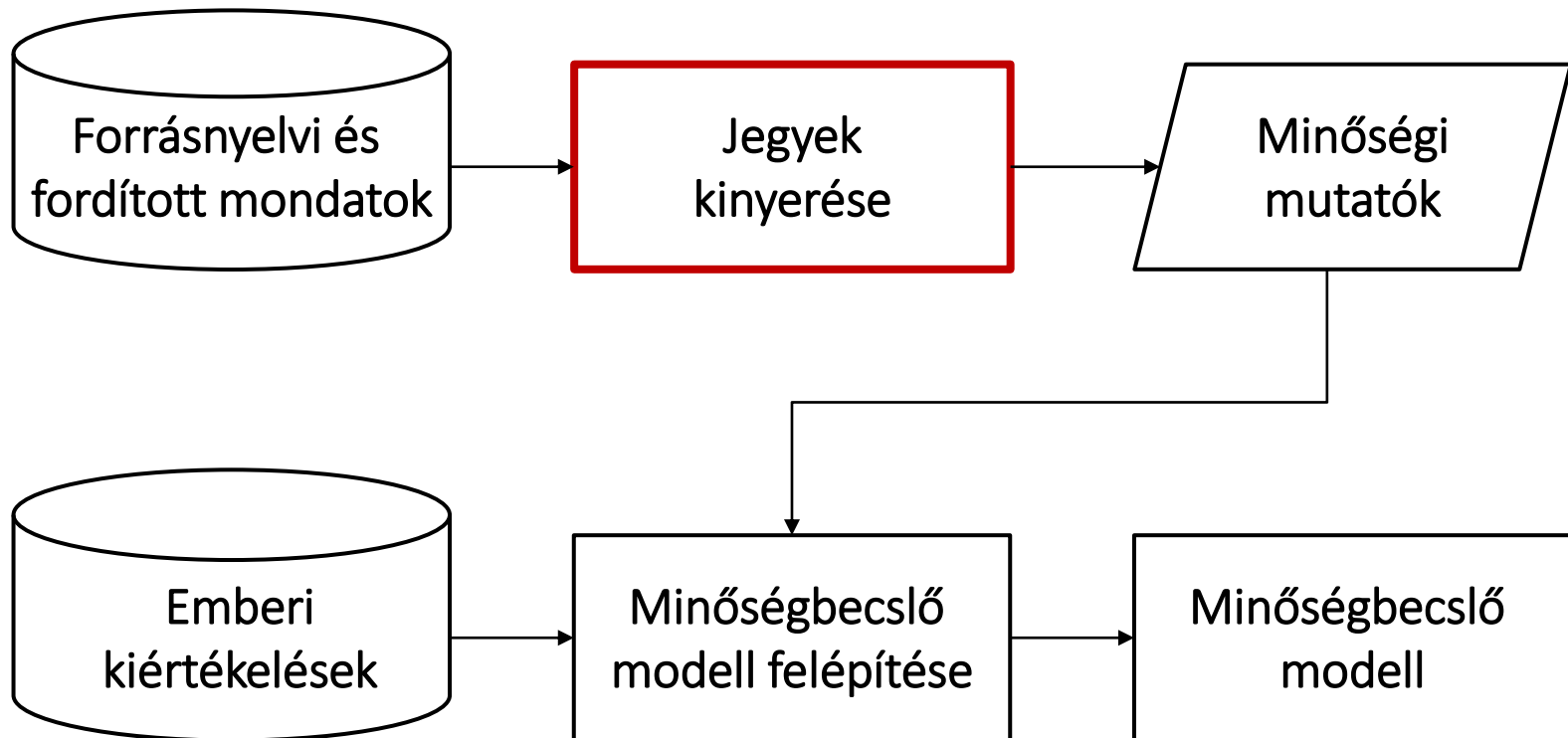
- Human-targeted Translation Edit/Error Rate
- Emberek a gépi fordító által lefordított mondatot minimális számú lépéssel kijavítják úgy, hogy a jelentés megegyezzen a referenciamondatével.
- Az így létrejött új referenciamondatra számolunk TER-t.

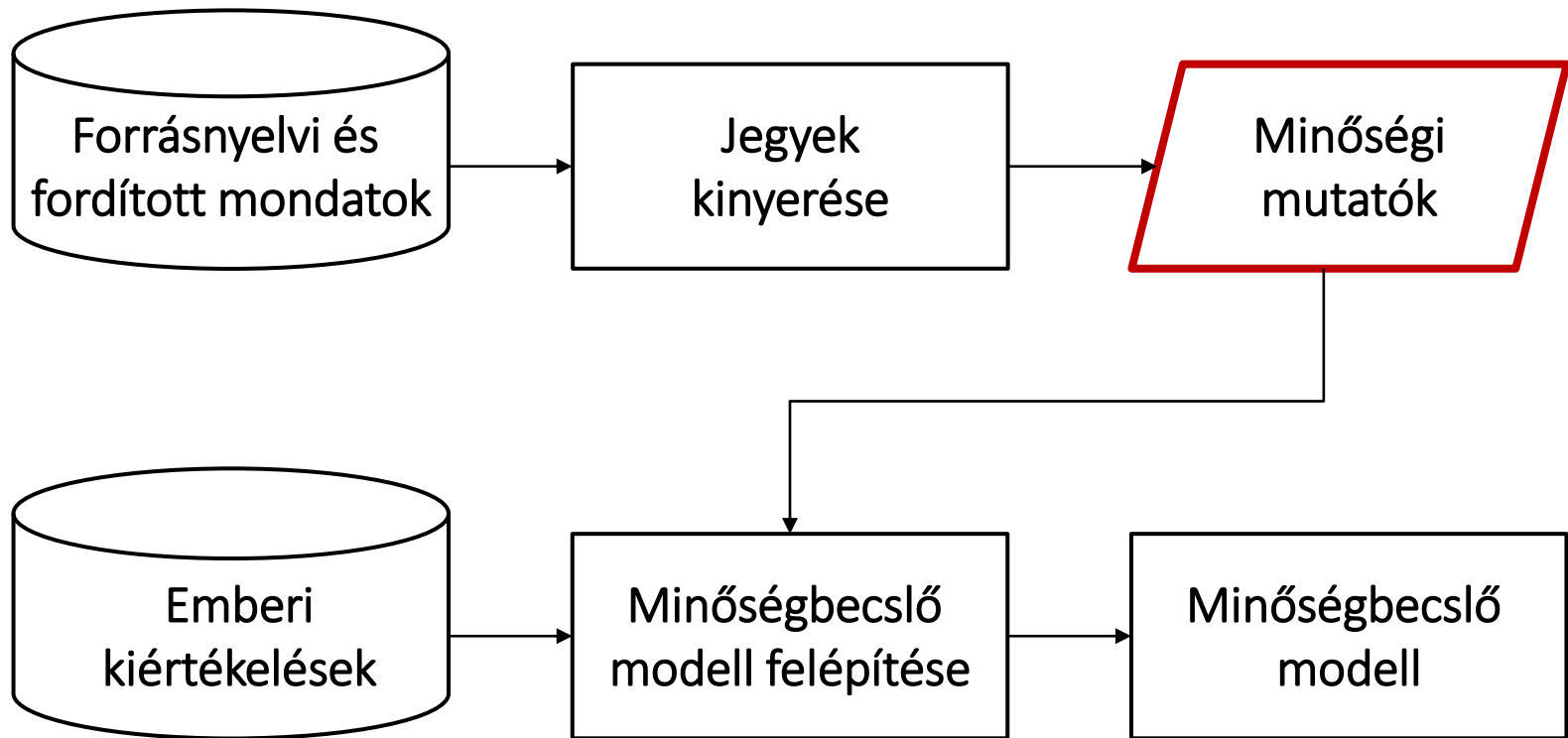


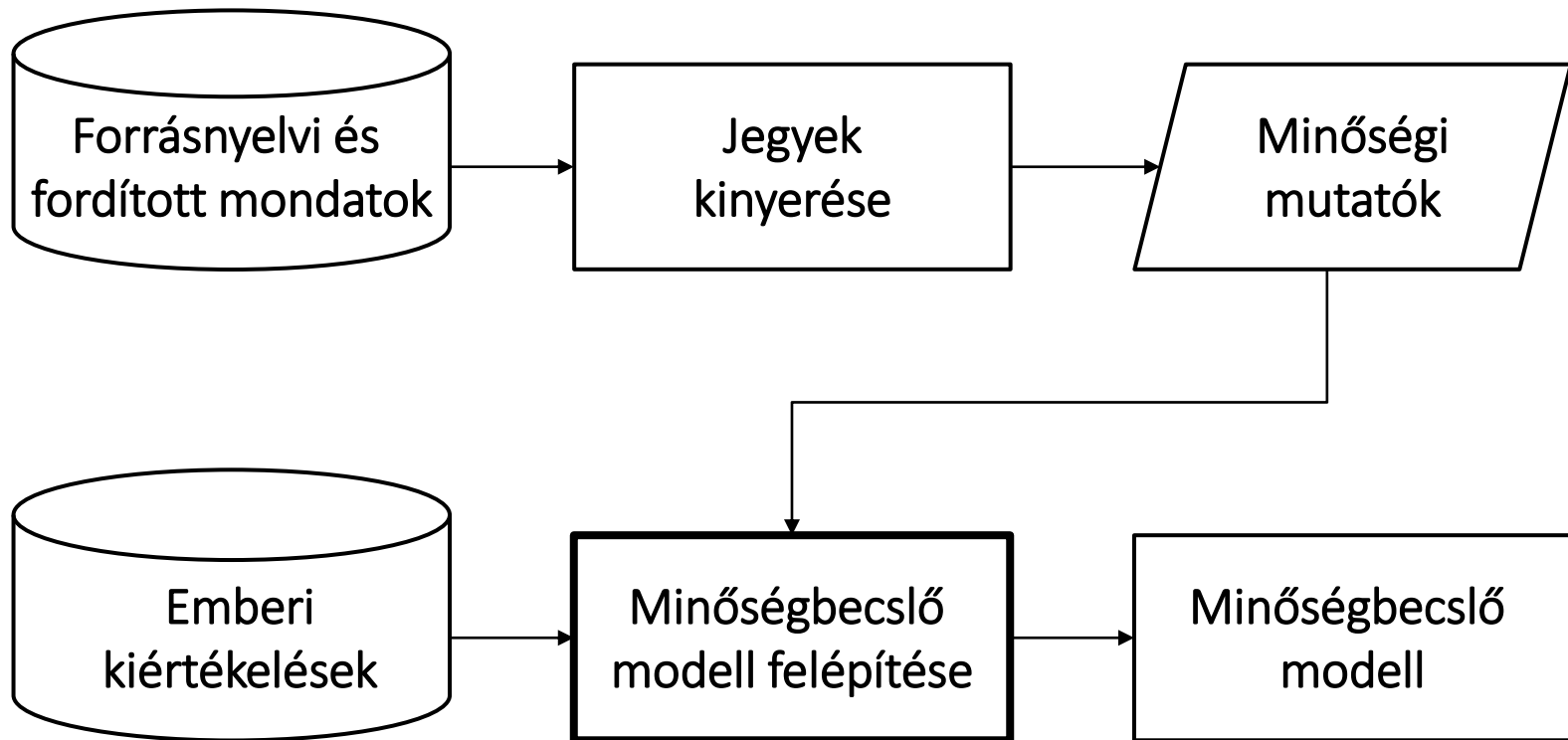
- Referenciafordítás nélküli módszer
- Valós időben tud kiértékelni
- Gépi tanuláson alapuló minőségbecslő modell
 - Regressziós / osztályzási feladat

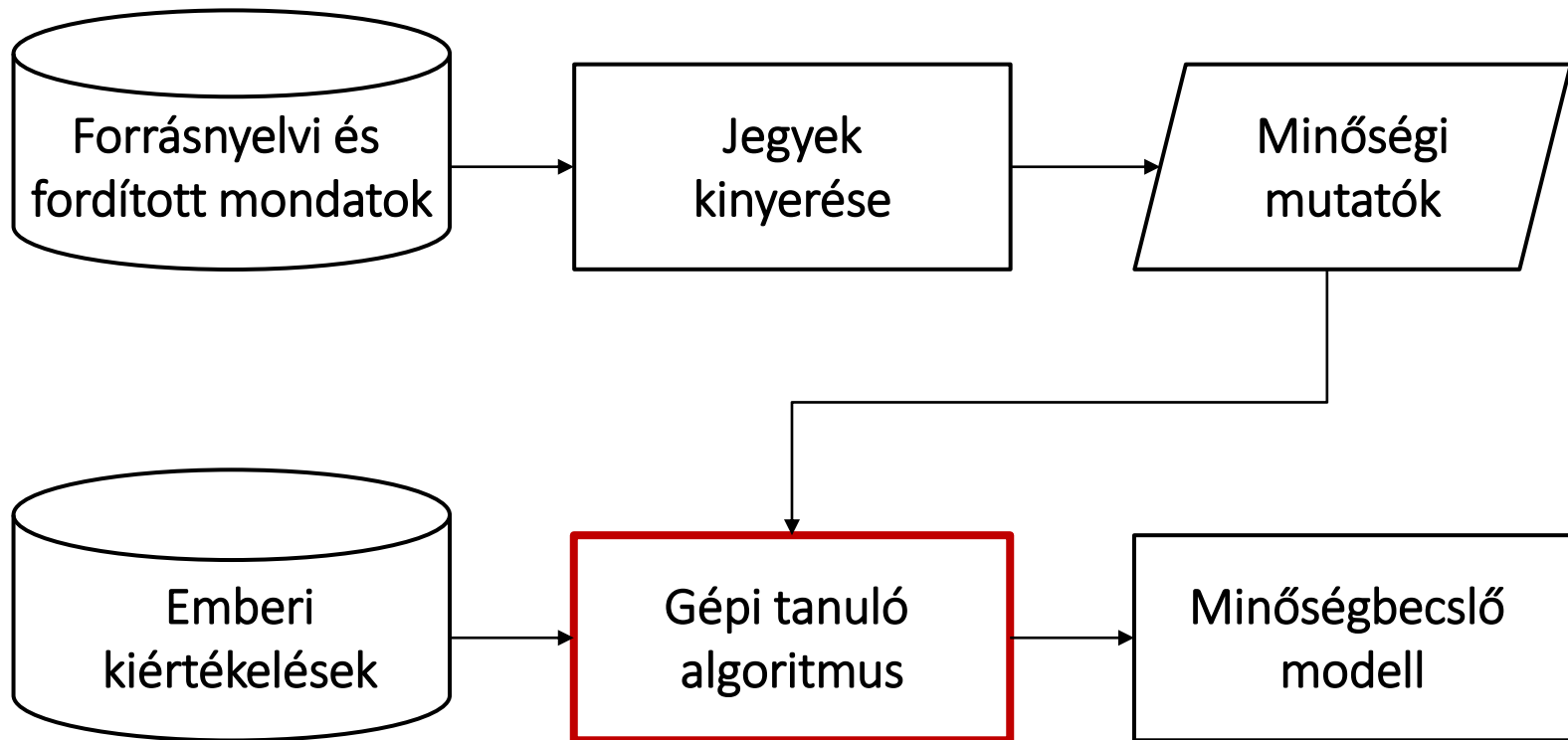


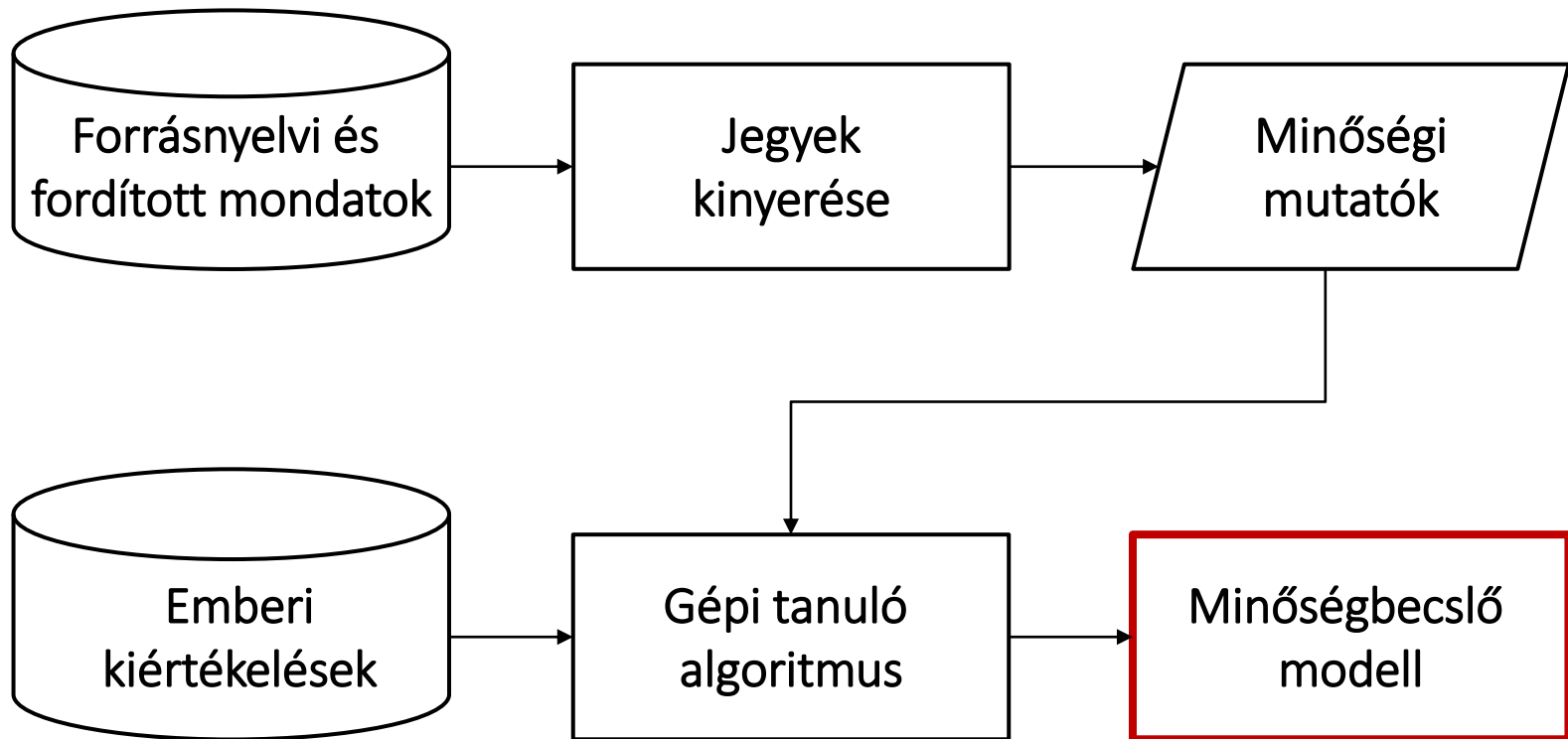


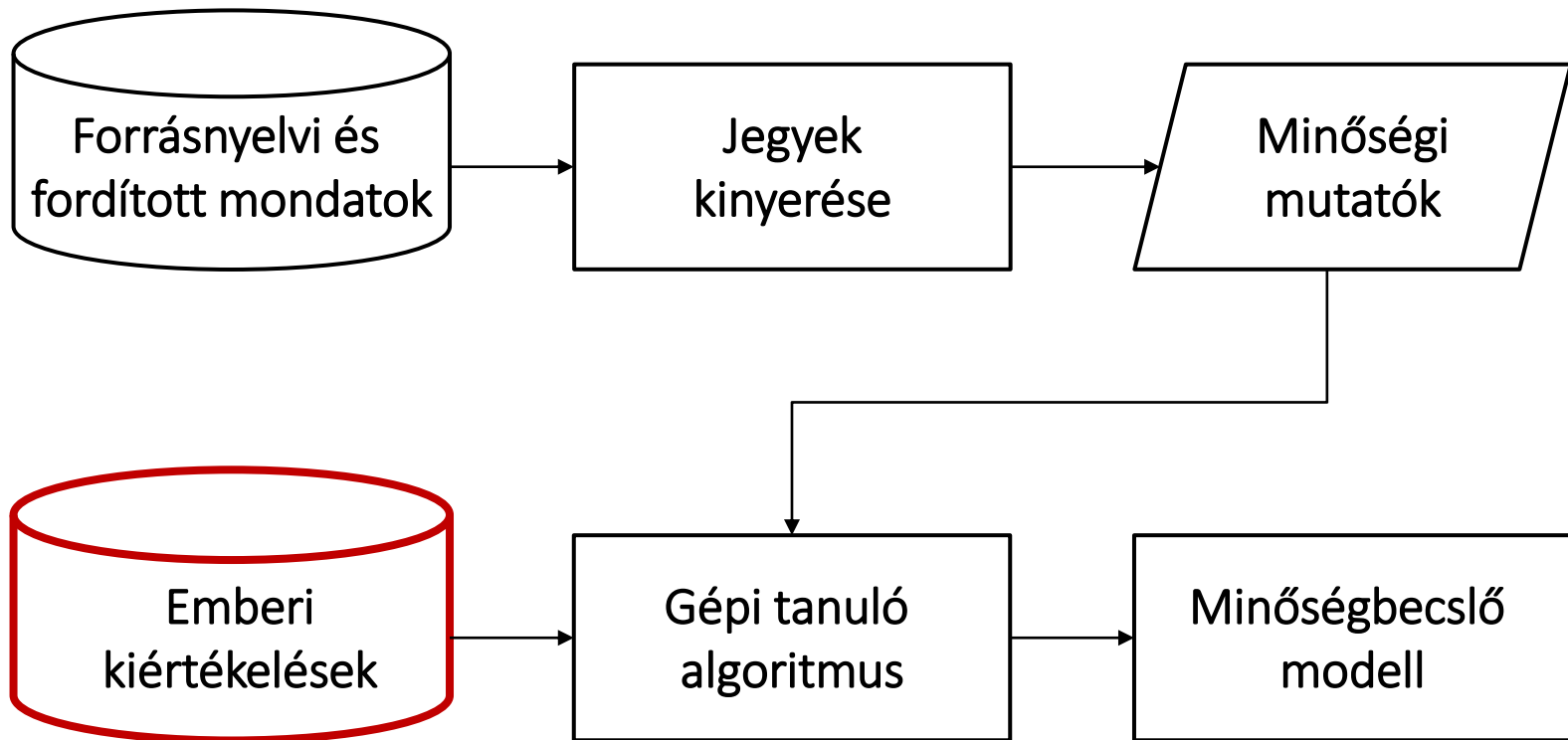








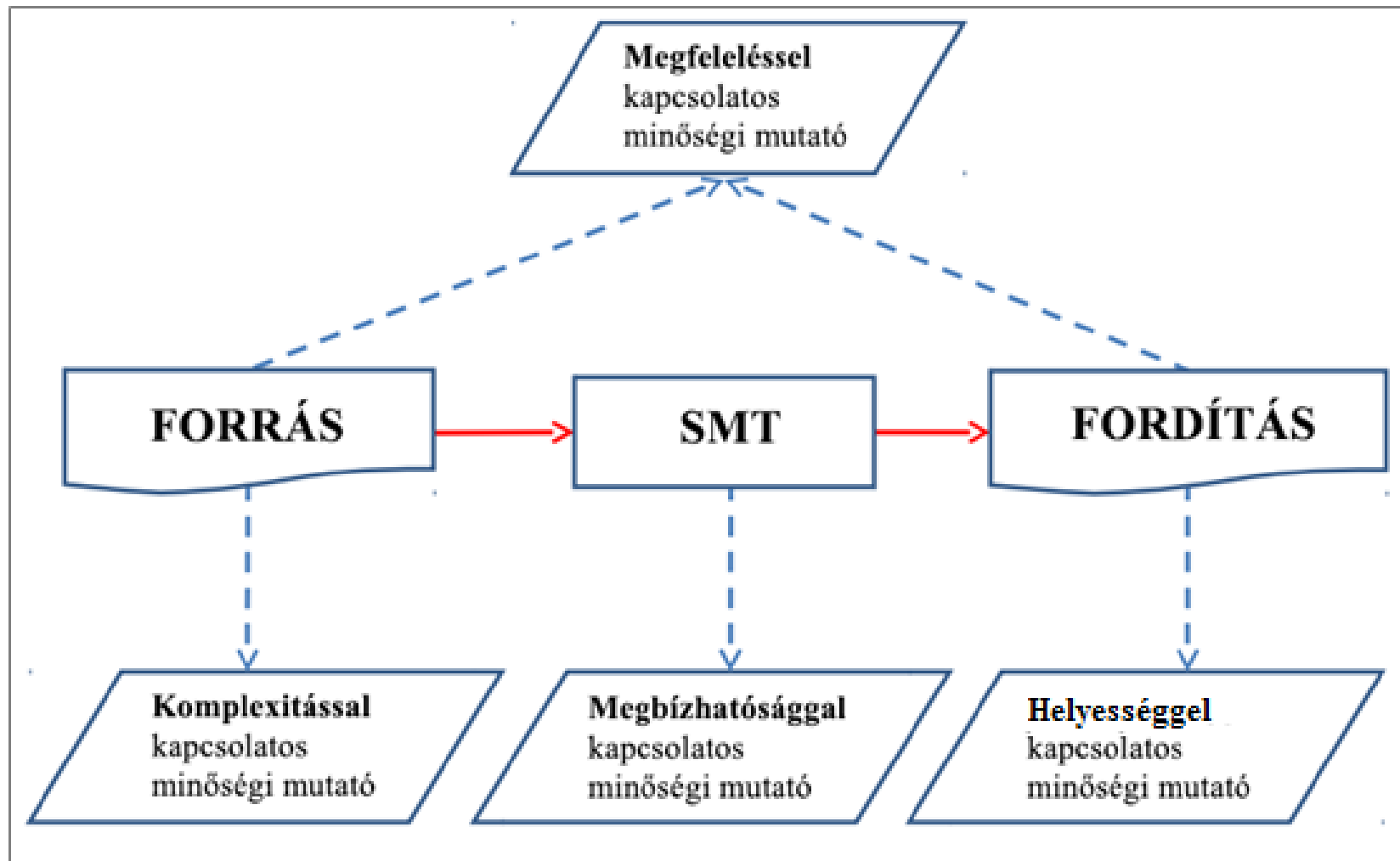






Forrásnyelvi mondat	Célnyelvi (lefordított) mondat	Minőségi mutató						Becslés	Emberi kiértékelés
She shot herself.	Lelőtte magát.	4.0	3.0	1.33	0.75	0.25	=	?	3.943
Now, as i have never known any ...	Most, ahogy még sohasem ismert ...	26.0	19.0	1.37	0.73	0.26	=	?	2.914
No work, no dinner.	Nincs munka, nincs vacsora.	8.0	8.0	1.00	1.00	0.00	=	?	3.596
But there seemed no place to go ...	De nem volt hová mennem ...	8.0	6.0	1.33	0.75	0.25	=	?	2.500
We tramped through the dim corridors ...	Mi órákon át a sötét folyosókon, kanyargós ...	34.0	29.0	1.17	0.85	0.14	=	?	2.627
Then he jumped like a cat ...	Aztán felugrott, mint egy macska, ...	28.0	26.0	1.08	0.93	0.07	=	?	3.275

Jegyek kinyerése





Jegyek

Igék aránya a forrásmondatban.

Igék aránya a célmondatban.



Jegyek

Igék aránya a forrásmondatban.

Igék aránya a célmondatban.

Főnevek aránya a forrásmondatban.

Főnevek aránya a célmondatban.

Melléknevek aránya a forrásmondatban.

Melléknevek aránya a célmondatban.

Igék aránya a forrásmondatban.

Igék aránya a célmondatban.

Főnevek aránya a forrásmondatban.

Főnevek aránya a célmondatban.

Melléknevek aránya a forrásmondatban.

Melléknevek aránya a célmondatban.

A forrás- és a célnyelvi mondatban a tokenek aránya.

Célmondatban a tokenek száma / forrásmondatban a tokenek száma.

Tokenek száma és forrás és forrás hossz által normalizált cél abszolút értékben vett különbsége.

Igék aránya a forrásmondatban.

Igék aránya a célmondatban.

Főnevek aránya a forrásmondatban.

Főnevek aránya a célmondatban.

Melléknevek aránya a forrásmondatban.

Melléknevek aránya a célmondatban.

A forrás- és a célnyelvi mondatban a tokenek aránya.

Célmondatban a tokenek száma / forrásmondatban a tokenek száma.

Tokenek száma és forrás és forrás hossz által normalizált cél abszolút értékben vett különbsége.

Hibás zárójelek száma.

Hibás idézőjelek száma.

Igék aránya a forrásmondatban.

Igék aránya a célmondatban.

Főnevek aránya a forrásmondatban.

Főnevek aránya a célmondatban.

Melléknevek aránya a forrásmondatban.

Melléknevek aránya a célmondatban.

A forrás- és a célnyelvi mondatban a tokenek aránya.

Célmondatban a tokenek száma / forrásmondatban a tokenek száma.

Tokenek száma és forrás és forrás hossz által normalizált cél abszolút értékben vett különbsége.

Hibás zárójelek száma.

Hibás idézőjelek száma.

Számok aránya százalékban, a forrásmondatban.

Számok aránya százalékban, a célmondatban.

A forrás- és a célmondatban lévő számok számosságának különbsége abszolút értékben, forrásmondat hosszával normalizálva.

Tokenek száma a forrásmondatban, amelyek nem csak a-z betűt tartalmaznak.

Tokenek százaléka a célmondatban, amelyek nem csak a-z betűt tartalmazzák.

Forrásmondat perplexitása.

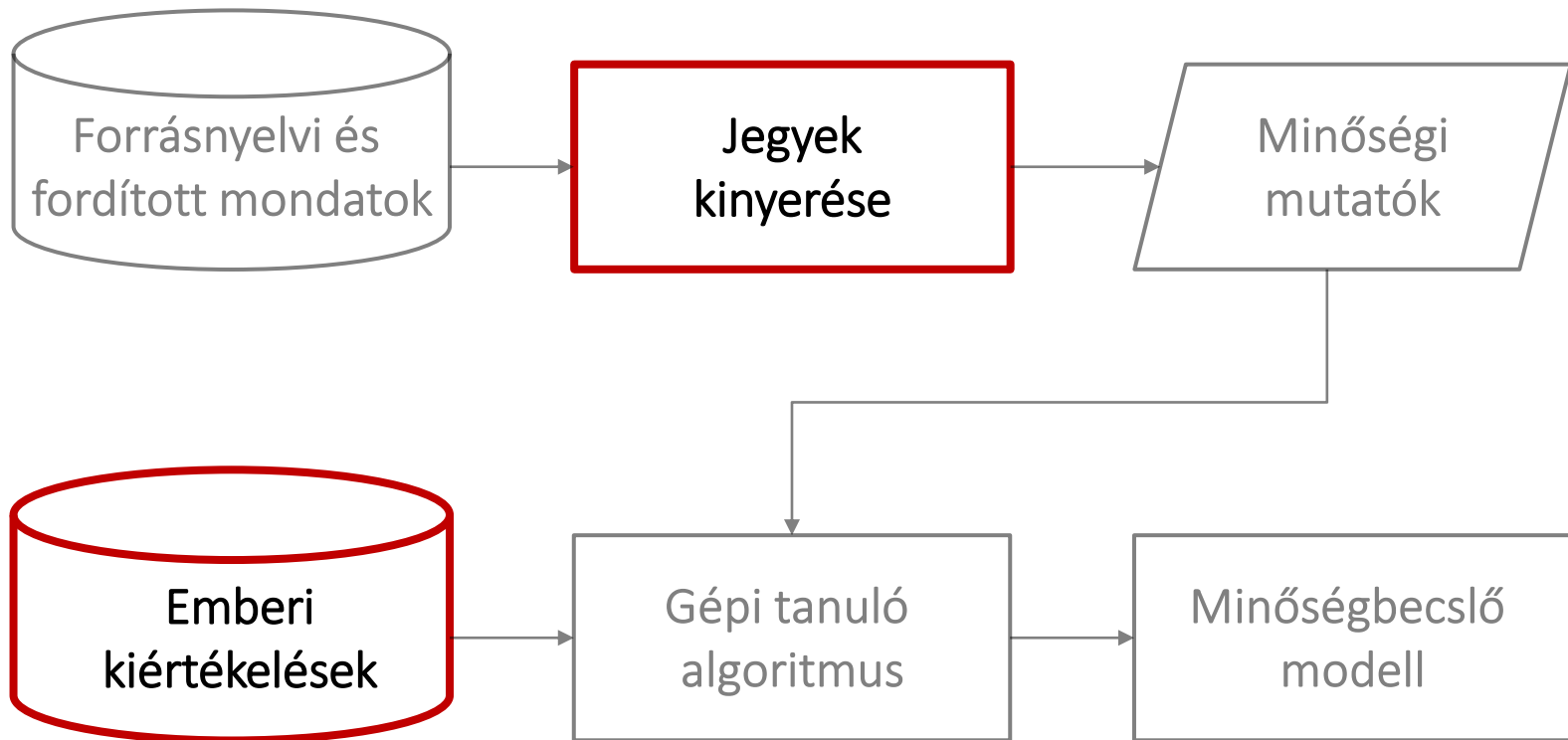
Forrásmondat perplexitása mondatvégi jelzés nélkül.

Célmondat perplexitása.

célmondat perplexitása mondatvégi jelzés nélkül.

Igék	Átlagos száma minden forrásszó fordításának a mondatban (giza1 küszöb: valószínűség > 0,01).
Igék	Átlagos száma minden forrásszó fordításának a mondatban (giza1 küszöb: valószínűség > 0,05).
Főne	Átlagos száma minden forrásszó fordításának a mondatban (giza1 küszöb: valószínűség > 0,1).
Főne	Átlagos száma minden forrásszó fordításának a mondatban (giza1 küszöb: valószínűség > 0,5).
Mellk	Átlagos száma minden forrásszó fordításának a mondatban (giza1 küszöb: valószínűség > 0,01).
Mellk	forráskorpuszban lévő minden szó gyakoriságával súlyozva.
A fo	Átlagos száma minden forrásszó fordításának a mondatban (giza1 küszöb: valószínűség > 0,5).
Célr	forráskorpuszban lévő minden szó gyakoriságával súlyozva.
Token	Átlagos száma minden forrásszó fordításának a mondatban (giza1 küszöb: valószínűség > 0,1).
Hibá	forráskorpuszban lévő minden szó gyakoriságával súlyozva.
Hibá	Átlagos száma minden forrásszó fordításának a mondatban (giza1 küszöb: valószínűség > 0,2).
Szán	forráskorpuszban lévő minden szó gyakoriságával súlyozva.
Szán	Átlagos száma minden forrásszó fordításának a mondatban (giza1 küszöb: valószínűség > 0,5).
A fo	forráskorpuszban lévő minden szó inverz gyakoriságával súlyozva.
Token	Átlagos száma minden forrásszó fordításának a mondatban (giza1 küszöb: valószínűség > 0,1).
Token	forráskorpuszban lévő minden szó inverz gyakoriságával súlyozva.
Forrá	Átlagos unigram gyakoriság a második kvartilisben lévő gyakorisága (kis gyakoriságú szavak) a forrásnyelvi korpuszban.
Forrá	Átlagos unigram gyakoriság a harmadik kvartilisben lévő gyakorisága (kis gyakoriságú szavak) a forrásnyelvi korpuszban.
Forrá	Átlagos bigram gyakoriság a második kvartilisben lévő gyakoriság (kis gyakoriságú szavak) a forrásnyelvi korpuszban.
Célr	Átlagos bigram gyakoriság a harmadik kvartilisben lévő gyakorisága (kis gyakoriságú szavak) a forrásnyelvi korpuszban.
célm	Átlagos trigram gyakoriság a második kvartilisben lévő gyakorisága (kis gyakoriságú szavak) a forrásnyelvi korpuszban.
célm	Átlagos trigram gyakoriság a harmadik kvartilisben lévő gyakorisága (kis gyakoriságú szavak) a forrásnyelvi korpuszban.
célm	A korpuszban előforduló különböző bigramok százaléka (minden kvartilisben).
célm	A korpuszban előforduló különböző trigramok százaléka (minden kvartilisben).

Igék	Átlagos s	A forrásmondatban és a célmondatban lévő vesszők számának különbsége abszolút értékben.
Igék	Átlagos s	A forrásmondatban és a célmondatban lévő vesszők számának különbsége abszolút értékben, célmondat
	Átlagos s	hosszával normalizálva.
Főne	Átlagos s	A forrásmondatban és a célmondatban lévő kettőspontok számának különbsége abszolút értékben.
Főne	Átlagos s	A forrásmondatban és a célmondatban lévő kettőspontok számának különbsége abszolút értékben, célmondat
	forráskoi	hosszával normalizálva.
Mell	Átlagos s	A forrásmondatban és a célmondatban lévő pontosvesszők számának különbsége abszolút értékben.
Mell	forráskoi	A forrásmondatban és a célmondatban lévő pontosvesszők számának különbsége abszolút értékben,
	Átlagos s	célmondat hosszával normalizálva.
A foi	forráskoi	A forrásmondatban és a célmondatban lévő kérdőjelek számának különbsége abszolút értékben.
Célmr	Átlagos s	A forrásmondatban és a célmondatban lévő kérdőjelek számának különbsége abszolút értékben, célmondat
	forráskoi	hosszával normalizálva.
Toke	Átlagos s	A forrásmondatban és a célmondatban lévő felkiáltójelek számának különbsége abszolút értékben.
Hibá	forráskoi	A forrásmondatban és a célmondatban lévő felkiáltójelek számának különbsége abszolút értékben, célmondat
	Átlagos s	hosszával normalizálva.
Hibá	forráskoi	A forrás- és a célmondatban lévő írásjegyek számának különbsége abszolút értékben, célmondat hosszával
Szán	Átlagos s	normalizálva.
	forráskoi	A célszó előfordulásának száma a célhipotézisben.
Szán	Átlagos s	Fordítások átlaga minden forrásszóra a mondatban (IBM 1 táblázatból kinyerhető adat, ahol $p(t s) > 0,2$).
A foi	forráskoi	Fordítások átlaga minden forrásszóra a mondatban, súlyozva a forrásnyelvi korpuszban lévő minden szó inverz
	Átlagos s	gyakoriságával (IBM 1 táblázatból kinyerhető adat, ahol $p(t s) > 0,01$).
Toke	forráskoi	Forrásnyelvi korpuszban lévő negyedik kvartilisben lévő forrásszó unigramjának gyakorisága (kis gyakoriságú
	Átlagos t	szavak) százalékban.
Toke	Átlagos t	Forrásnyelvi korpuszban lévő negyedik kvartilisben lévő forrásszó unigramjának gyakorisága (nagy gyakoriságú
	Átlagos k	szavak) százalékban.
Forrá	Átlagos k	Forrásnyelvi korpuszban lévő első kvartilisben lévő forrásszó bigramjának gyakorisága százalékban.
Forrá	Átlagos t	Forrásnyelvi korpuszban lévő negyedik kvartilisben lévő forrásszó bigramjának gyakorisága százalékban.
Célmr	Átlagos t	Forrásnyelvi korpuszban lévő első kvartilisben lévő forrásszó trigramjának gyakorisága százalékban.
célm	A korpus	Forrásnyelvi korpuszban lévő negyedik kvartilisben lévő forrásszó trigramjának gyakorisága százalékban.
	A korpus	Forrásmondat unigramjainak előfordulási százalékai a korpuszban (SMT tanító korpusz alapján)..





- Gépi fordítás célja: megértés, információkinyerés, NEM a tökéletes fordítás
- Gépi fordítás kiértékelése:
 - Minden esetben az emberi kiértékelésen / fordításon alapszik
 - Referenciafordítással történő kiértékelés
 - Minőségbecslés
 - Nem használ referenciafordítást.
 - Egyik módszer sem éri el az elvárt eredményt.



Köszönöm a figyelmet



ध्यान देने के लिए आपका धन्यवाद
आप अपना ध्यान के लिए धन्यवाद
थैंक यू फॉर योर अटेंशन

Thank you for your attention
Thank you for your attention
Thank you for the attention

Merci de votre attention
Je vous remercie pour votre attention
Merci pour l'attention

感谢您的关注
谢谢你的关注
谢谢您的注意