Zainkó Csaba

Deep Learning a gyakorlatban Python és LUA alapon

NLP

SmartLab
Intelligent Interactions

http://smartlab.tmit.bme.hu

NVIDIA
GPU
EDUCATION
CENTER

Jogi nyilatkozat

Jelen előadás diái a „*Deep Learning a gyakorlatban Python és LUA alapon*" című tantárgyhoz készültek és letölthetők a http://smartlab.tmit.bme.hu honlapról.

A diák nem helyettesítik az előadáson való részvételt, csupán emlékeztetőül szolgálnak.

Az előadás diái a szerzői jog védelme alatt állnak. Az előadás diáinak vagy bármilyen részének újra felhasználása, terjesztése, megjelenítése csak a szerző írásbeli beleegyezése esetén megengedett. Ez alól kivétel, mely diákon külső forrás külön fel van tüntetve.

Nagy házi feladat - beszámoló

- Jelentkezzetek vizsgára és ott mutassátok be
- Megajánlott jegyért és aláírásért is

Prezentációs fóliák

- PPT vagy PDF a Github repoba feltöltve a vizsga előtti nap éjfélig.
  - Téma bemutatása
  - Korábbi munkák
  - Kidolgozott megoldás
  - Kiértékelés
  - Összefoglalás
- Lehet angolul és magyarul is.

Nagy házi feladat - beszámoló

Prezentációs videó

- Unlisted YouTube link, a GitHub repo README.MD-jébe írjátok bele a vizsga előtti nap éjfélig feltöltve.
- 10 perc (+/- 1 perc)
- A prezentációs fóliák bemutatása
- Egyik sarokban a előadó személy mutatása (webkamerával)
- Minden csapattag beszéljen

Javasolt software:

- Felvétel: OBS Studio
- Utómunkák: OpenShot
- vagy Teams meeting előzetes rögzítése

Kisházi

- 2. kijavítva
- 3. javítás alatt

NLP

Natural language processing

NLP- Natural language processing

- Több tématerület érintett:
  - Linguistics
  - Artificial intelligence
  - Computer science

- Cél:
  - Megérteni/feldolgozni az emberi nyelvet/beszédet

NLP alkalmazási területek

- Helyesírásellenőrzés
- Kulcsszókeresés
- Szinonimák
- Osztályozás, információ kinyerés,
- Gépi fordítás
- Dialógus rendszerek
- Beszédfelismerés
- Beszédszintézis
- Személyre/tartalomra szabott hirdetés
- Ügyfélszolgálati támogatás

Miért speciális?

- Nem formális nyelvek
  - Többértelmű szavak
  - Környezet függő jelentés
  - Multimodális információ továbbítás
    - Gesztus, hangsúly, dallam …
  - Folyamatos adatfolyam
    - Spontán beszédben mondat?

- Lassú kommunikáció -> extra modelleket használnak az emberek a megértésre.

Deep Learning Híradó

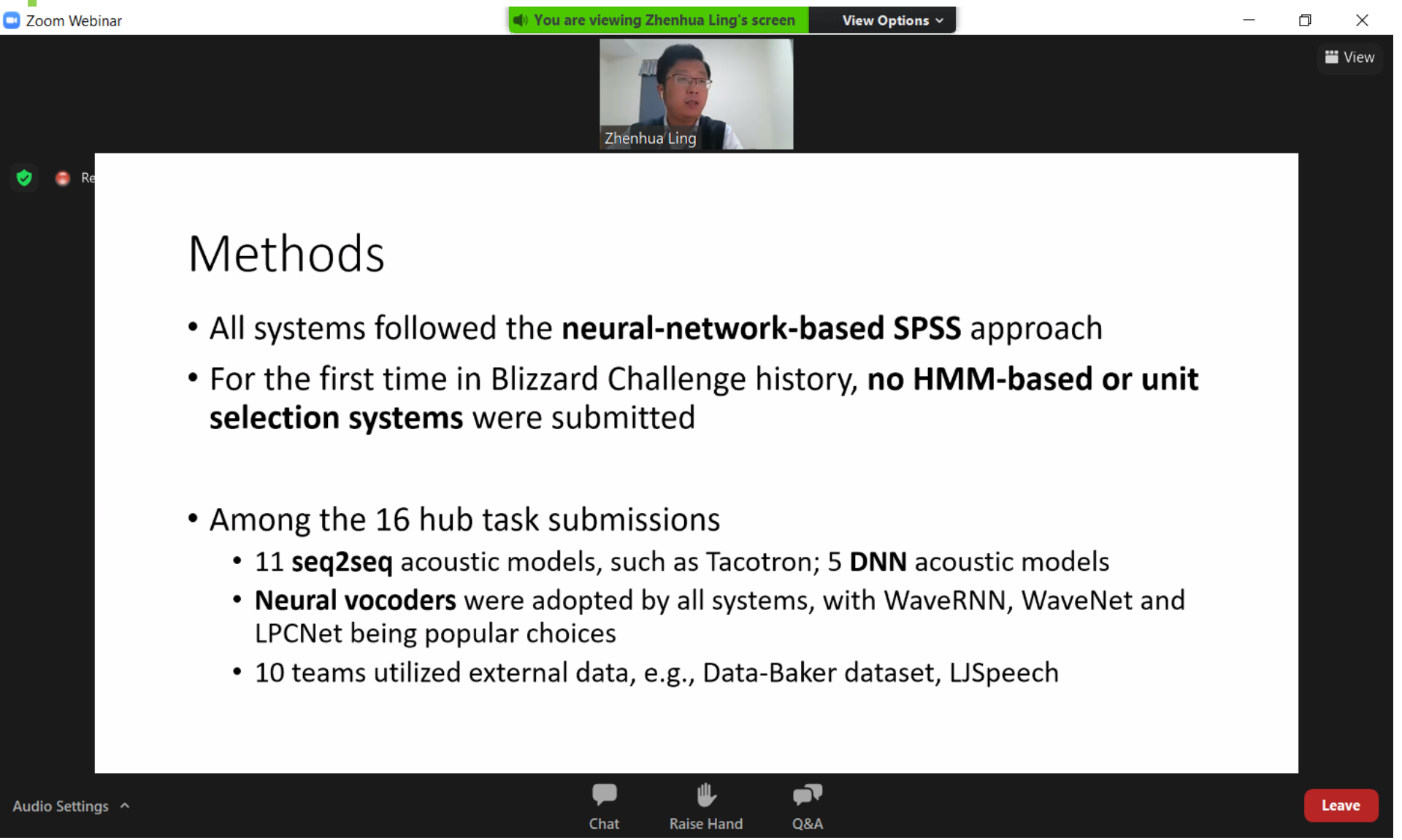
Hírek az elmúlt 168 órából

DNN mindenhol



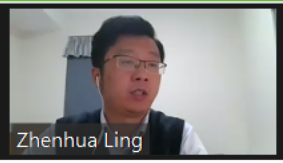
- Interspeech2020
  - https://www.isca-speech.org/archive/Interspeech_2020/
- The Blizzard Challenge 2020
  - https://www.isca-speech.org/archive/VCC_BC_2020/
- The Voice Conversion Challenge 2020
  - https://www.isca-speech.org/archive/VCC_BC_2020/

Blizzard Challenge



Blizzard Challenge

Zoom Webinar — You are viewing Zhenhua Ling's screen — View Options

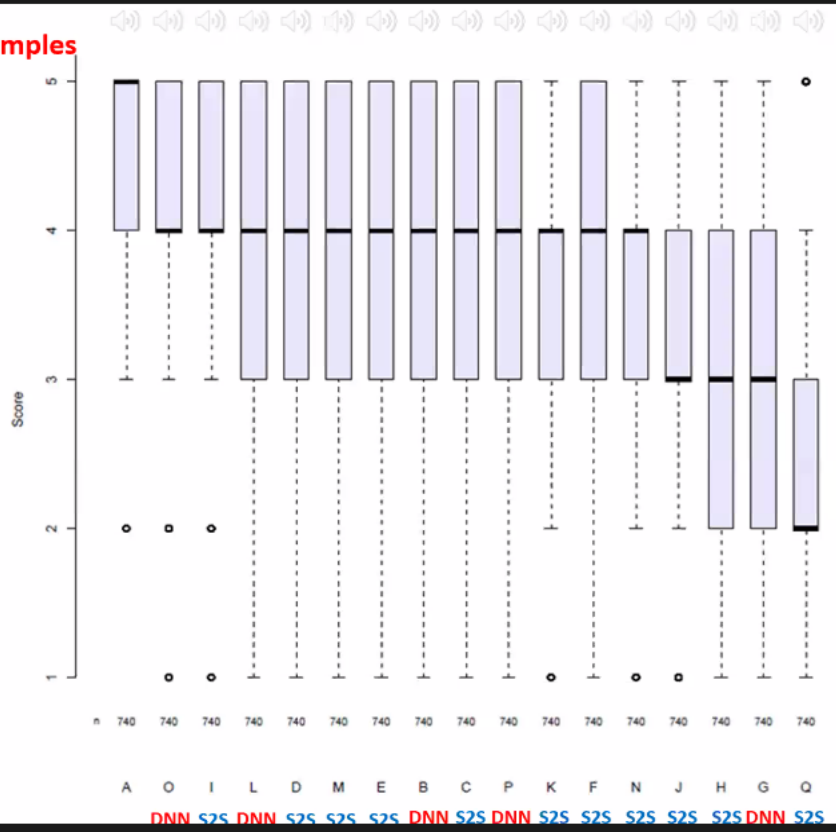


View

Task 2020-MH1 MOS (Naturalness)

News Samples

- **no system** was as natural as natural speech (MOS=4.7)
- **I** and **O** were significantly more natural than all other systems (MOS = 4.2)
- Techniques used by systems
  - **Acoustic model**
    - DNN (DNN), Seq2Seq (S2S)
  - **Vocoder**
    - WaveNet (WN), WaveRNN (WR), LPCNet (LN),
    - WaveGlow (WG), Parallel WaveGAN



Audio Settings — Chat — Raise Hand — Q&A — Leave

Hogyan kezdjük hozzá?

Nyelvi modellek

- Adott „mondat” valószínűsége (szósorozat)

„Szép idő van”

Egy mondat valószínűsége

„Szép idő van"

$$P(W) = P(|W| = 3, w_1 = szép, w_2 = idő, w_3 = van) =$$

$$P(w_1 = \text{szép} \mid w_0 = \text{<start>}) *$$
$$P(w_2 = idő \mid w_0 = \text{<start>}, w_1 = \text{szép}) *$$
$$P(w_3 = \text{van} \mid w_0 = \text{<start>}, w_1 = \text{szép}, w_2 = idő) *$$
$$P(w_4 = \text{<end>} \mid w_0 = \text{<start>}, w_1 = \text{szép}, w_2 = idő, w_3 = \text{van})$$

mj: $P(w_0 = \text{<start>}) = 1.0$

$$P(W) = \prod_{i=1}^{|W|+1} P(w_i | w_0 \cdots w_{i-1})$$

Maximum Likelihood becslés

$$P(W) = \prod_{i=1}^{|W|+1} P(w_i|w_0 \cdots w_{i-1})$$

$$P(w_i|w_0 \cdots w_{i-1}) = \frac{c(w_1 \cdots w_i)}{c(w_1 \cdots w_{i-1})}$$

<start> Szép idő lesz <end>
<start> Szép napos idő volt <end>

$$P(\text{idő}|\text{<start> szép}) = \frac{c(szép\ idő)}{c(szép)} = \frac{1}{2}$$

Uni-gram

$$P(W) = \prod_{i=1}^{|W|+1} P(w_i | w_0 \cdots w_{i-1})$$

- A környezet figyelmen kívül hagyása:

$$P(w_i | w_0 \cdots w_{i-1}) \approx P(w_i)$$

- Gond: P(Szép az idő ma.) = P(Az szép ma idő.)

Bi-gram

$$P(W) = \prod_{i=1}^{|W|+1} P(w_i | w_0 \cdots w_{i-1})$$

- A környezet figyelmen kívül hagyása:

$$P(w_i | w_0 \cdots w_{i-1}) \approx P(w_i | w_{i-1})$$

- Vagy tri-gram, …. stb

N-gram

$$P(W) = \prod_{i=1}^{|W|+1} P(w_i | w_0 \cdots w_{i-1})$$

$$P(w_i | w_0 \cdots w_{i-1}) \approx P\left(w_i \middle| w_{i-(N-1)}, w_{i-(N-1)+1,} \cdots w_{i-1}\right)$$

$$P(w_i | w_0 \cdots w_{i-1}) \approx P\left(w_i \middle| w_{i-N+1}, w_{i-N+2,} \cdots w_{i-1}\right)$$

Log Likelihood

$$P(W) = \prod_{i=1}^{|W|+1} P(w_i | w_0 \cdots w_{i-1})$$

pl.: $P(w_i | w_0 \cdots w_{i-1}) = 1.54 \cdot 10^{-20}$



szorzás

pl.: $\log P(w_i | w_0 \cdots w_{i-1}) = -19.81$



összeadás

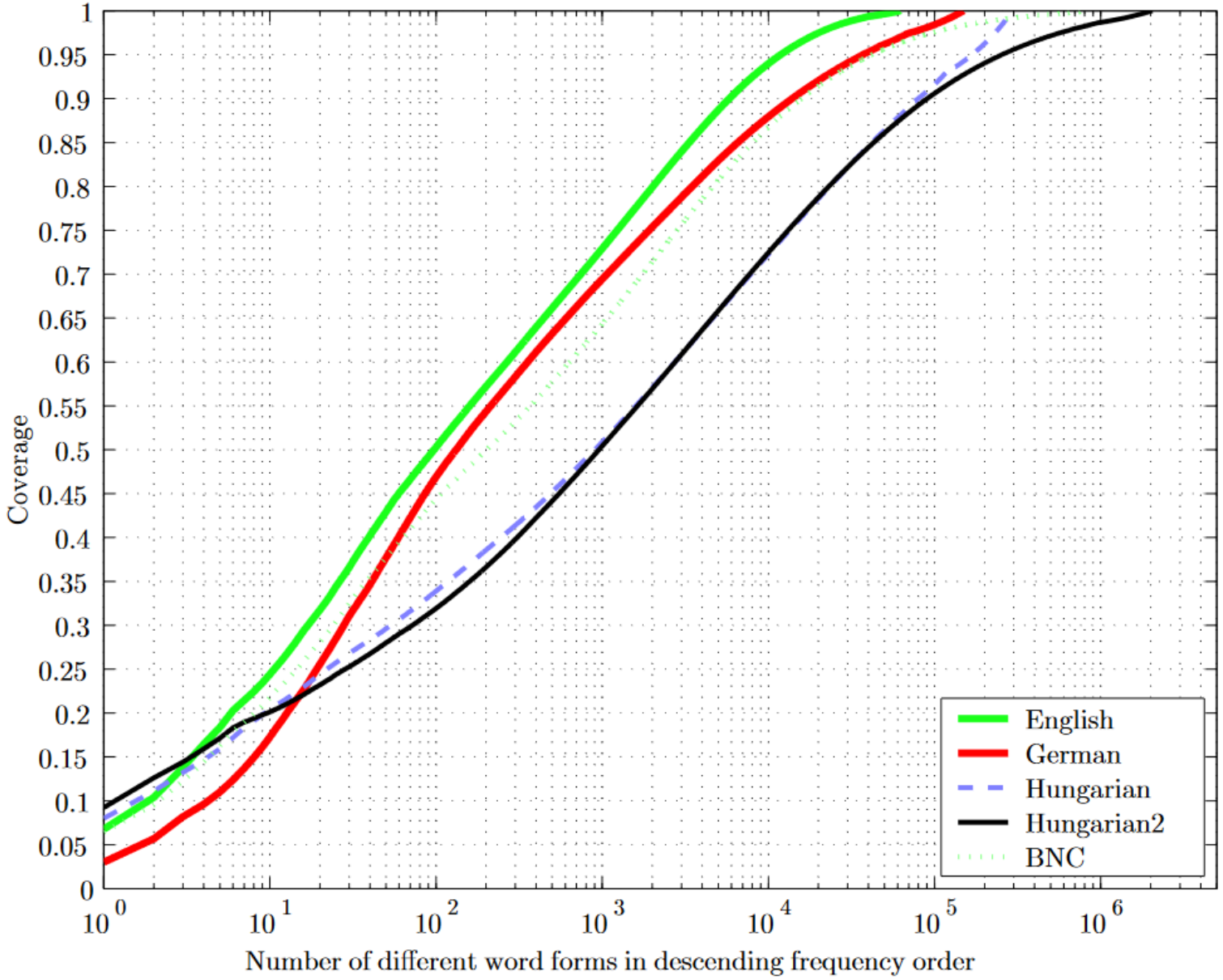
Log Likelihood

$$\mathrm{L}(W) = \prod_{i=1}^{|W|+1} P(w_i | w_0 \cdots w_{i-1})$$

$$\mathrm{l}(W) = \ln\left(\prod_{i=1}^{|W|+1} P(w_i | w_0 \cdots w_{i-1})\right) =$$

$$\mathrm{l}(W) = \sum_{i=1}^{|W|+1} \ln P(w_i | w_0 \cdots w_{i-1})$$

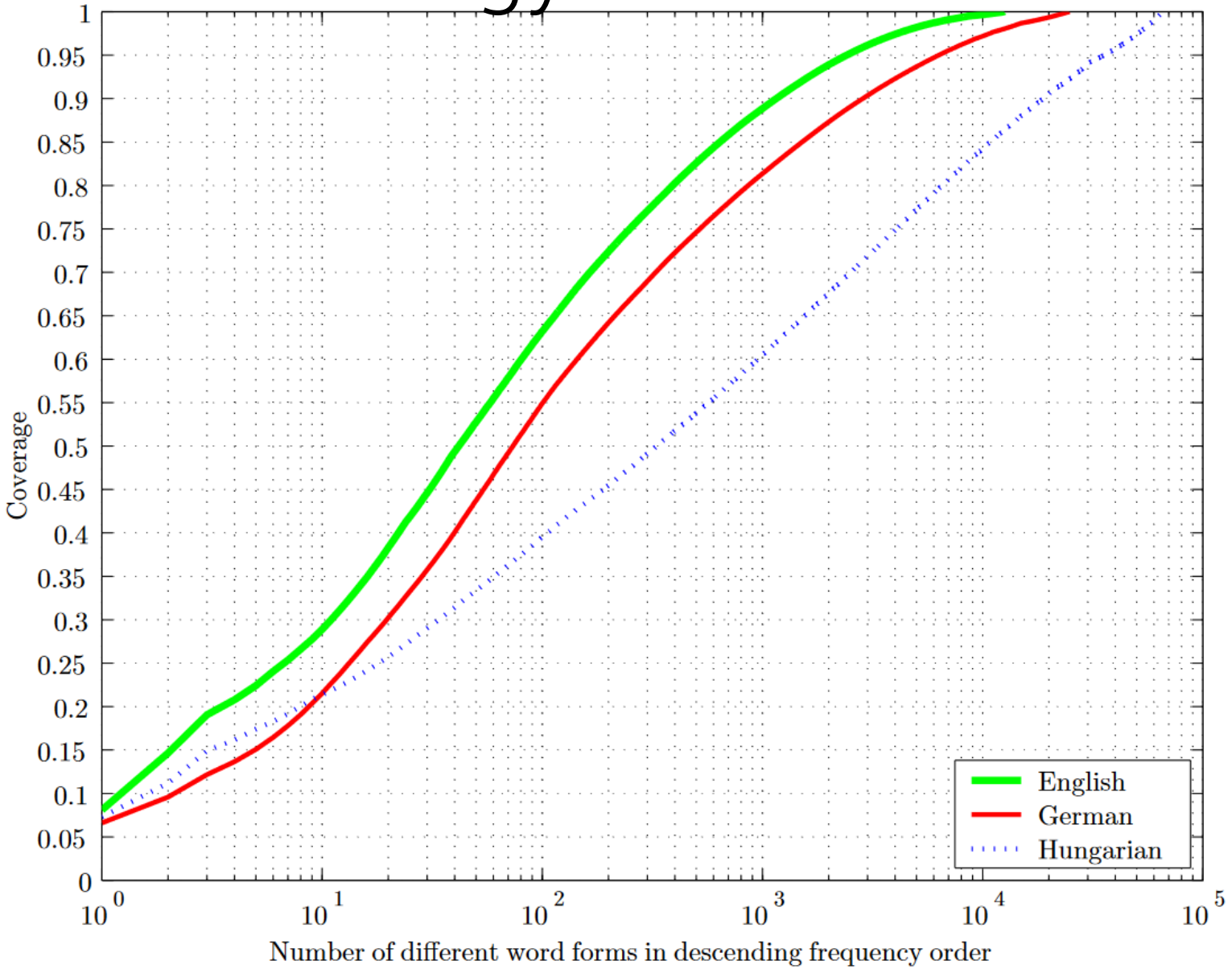
Szófedés


Coverage
Number of different word forms in descending frequency order
English
German
Hungarian
Hungarian2
BNC

Szófedés

- COV95 (3.5M, English) = 11 859
- COV95 (3.1M, German) = 36 982
- COV95 (2.5M, Hungarian) = 168 510

Szófedés – egy témakörön



Szavak kezelése

- One-hot

  Vektor, ahol 1 db 1-es van, a többi nulla

  alma: 0 0 0 0 0 0 0 0 0 0 0 1 0 0

- Vektor dimenziója: a szótár elemeinek szám: x millió?
- Hasonlóság: ortogonálisak

  kerékpár: 0 0 0 0 0 0 0 1 0 0 0 0 0 0

  bicikli: 0 0 0 0 0 1 0 0 0 0 0 0 0 0

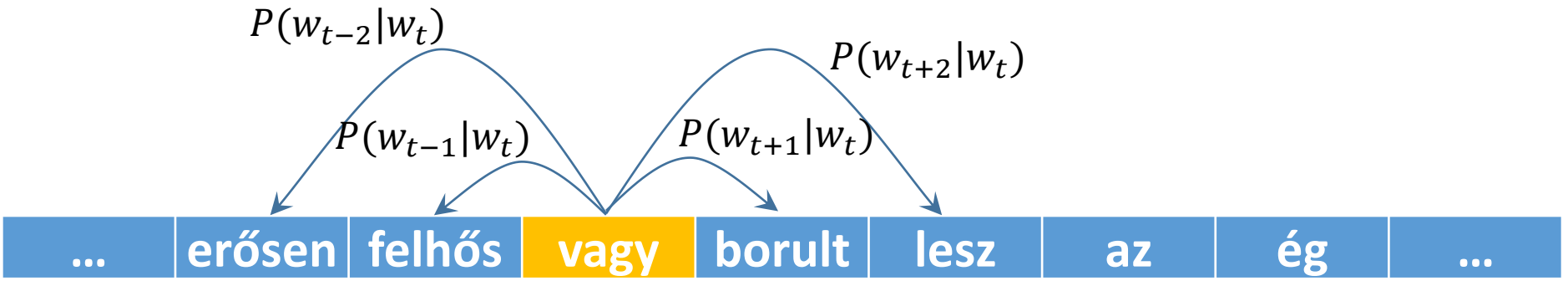
Hasonlóság?

- A szavak környezet jó megoldás lehet

- Minden szót egy vektorral jellemezzünk
- Jó sok szöveg
- Egy szó környezetét nézzük +-1,2,..N távolságig
- Használjuk a vektorok hasonlóságát, hogy a szavak valószínűségét meghatározzuk
- Addig módosítsuk a vektorokat, míg maximalizáljuk a valószínűségeket

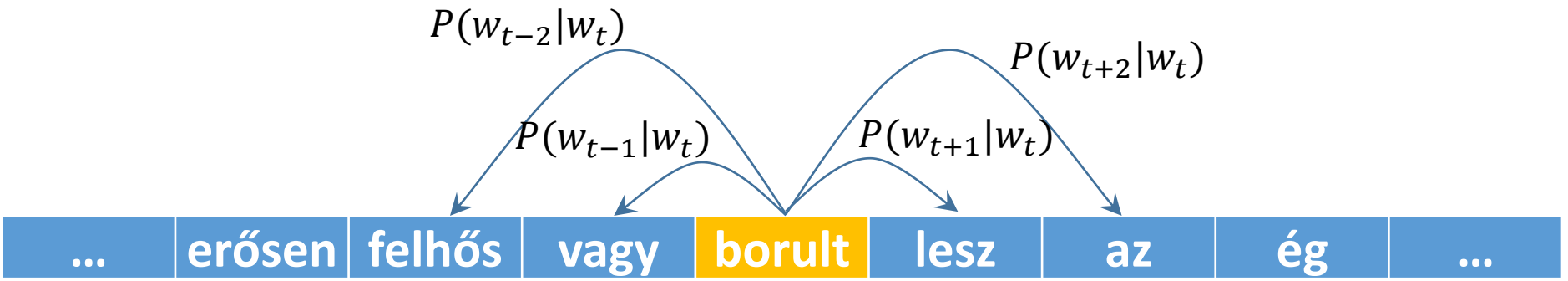
Word2Vec

„Többnyire erősen felhős vagy borult lesz az ég, hosszabb ideig tartó felszakadozás a Dunántúl déli felén lehet”



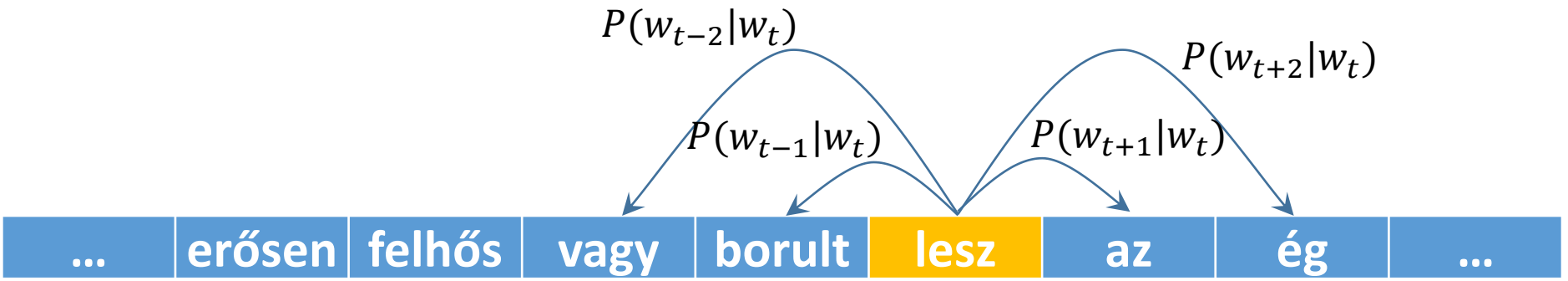
Word2Vec

„Többnyire erősen felhős vagy borult lesz az ég, hosszabb ideig tartó felszakadozás a Dunántúl déli felén lehet”



Word2Vec

„Többnyire erősen felhős vagy borult lesz az ég, hosszabb ideig tartó felszakadozás a Dunántúl déli felén lehet”



word2vec

Költség függvény

- Likelihood:

$$L(\theta) = \prod_{t=1}^{|szavak|} \prod_{\substack{-a \le k \le a \\ k \neq 0}} P(w_{t+k}|w_t;\theta)$$

- Költség:

$$J(\theta) = -\frac{1}{|szavak|} logL(\theta) =$$

$$= -\frac{1}{|szavak|} \sum_{t=1}^{|szavak|} \sum_{\substack{-a \le k \le a \\ k \neq 0}} logP(w_{t+k}|w_t;\theta)$$

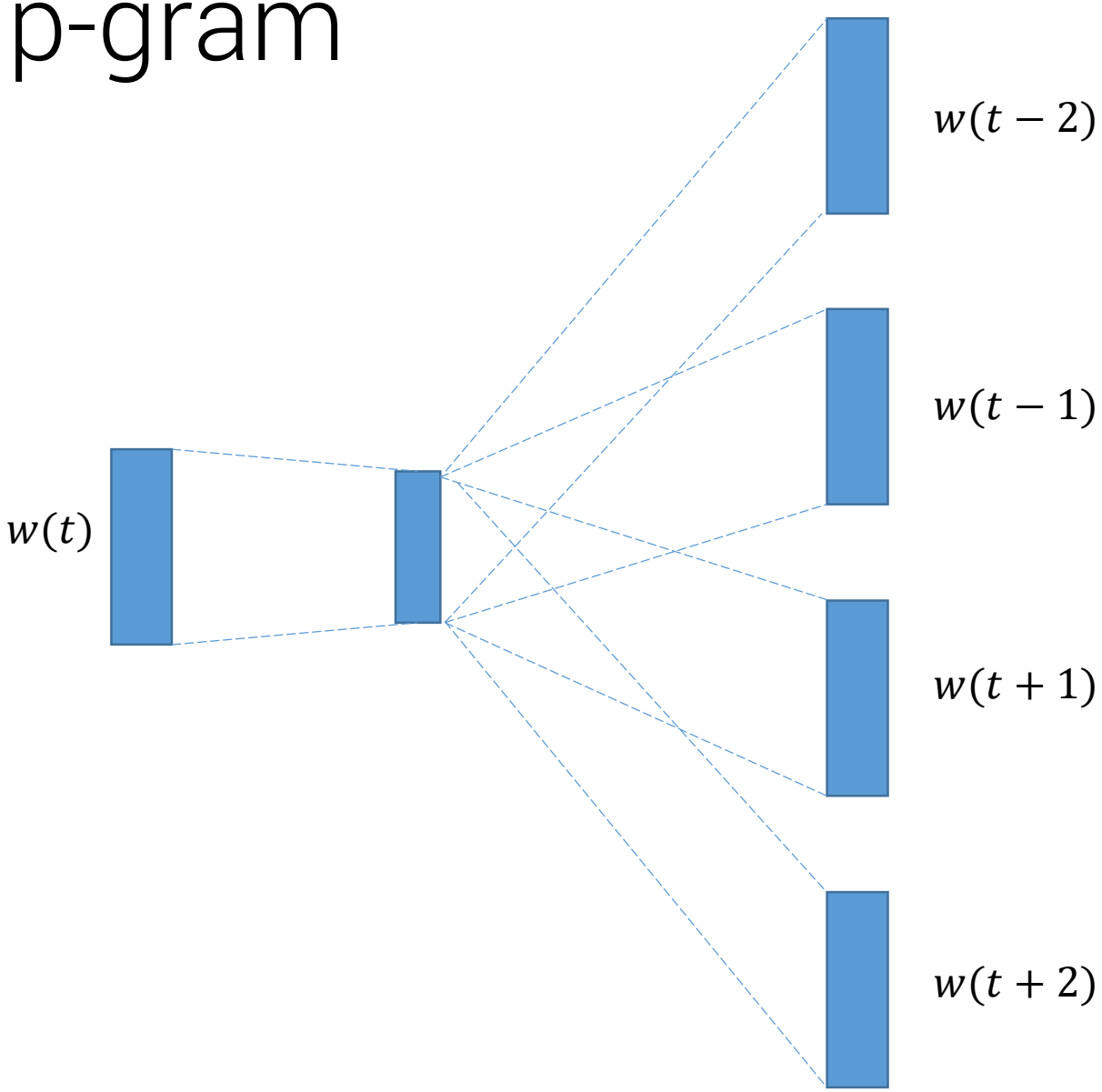
Költség függvény

$$P(w_{t+k}|w_t;\theta) \qquad P(o|c)$$

- Középső szó: v_w
- Környezeti szó: u_w

$$P(o|c) = \frac{e^{u_o^T v_c}}{\sum_{w \in V} e^{u_w^T v_c}}$$

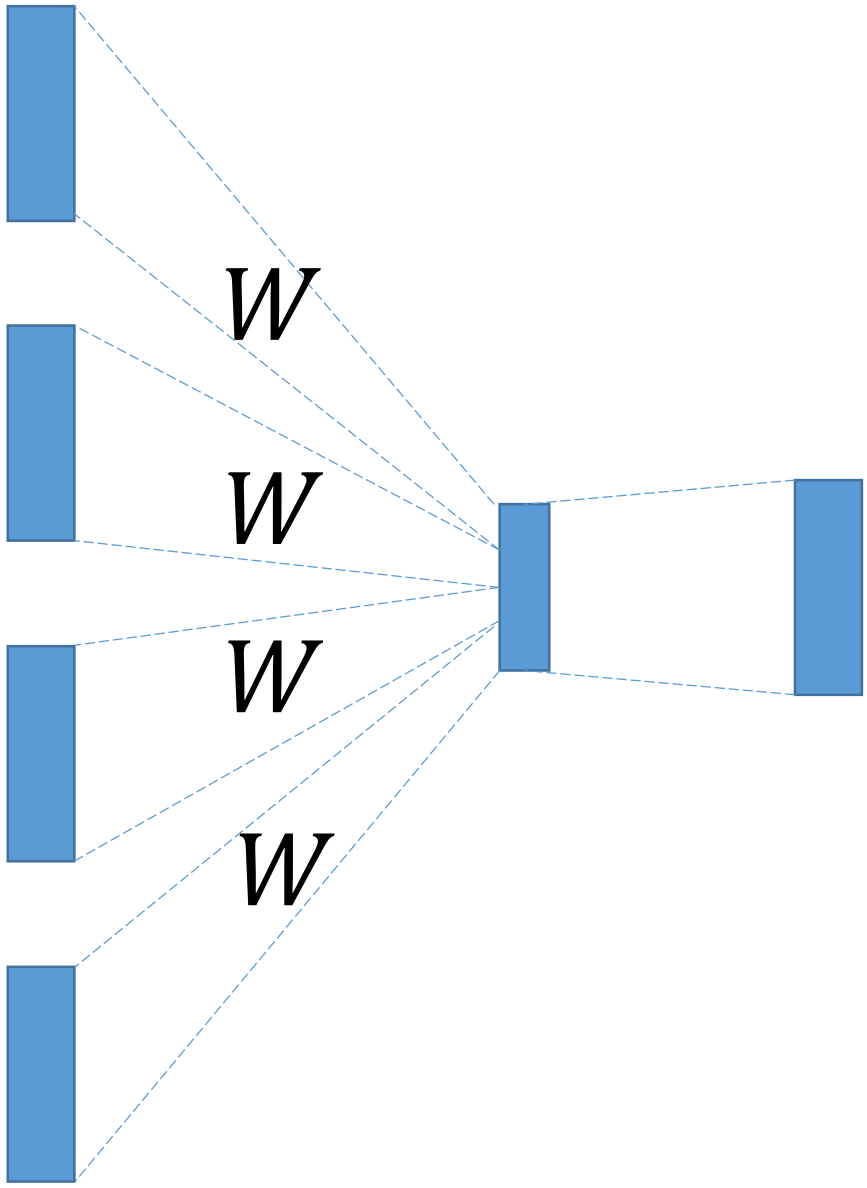
Skip-gram



Költség függvény

$$J_t(\theta) = \log \sigma \left(u_o^T v_c\right) + \sum_{j \sim P(w)} \left[\log \sigma \left(-u_j^T v_c\right)\right]$$

N-gram



Zsák

- Eredetileg NLP (természtesnyelv-feldolgozás) és IR (információkinyerés), ma már általánosított objektumleírás

- A vektor minden eleme egy-egy jellemzőnek felel meg
- Benne van / nincs benne - I(.) – logikai (boolean)
- Hányszor van benne – C(.) – egész (integer)

- Analógia: hisztogram

Zsák

- Gyümölcsért megyünk a piacra. Kapható alma, körte, barack, eper, meggy és kivi.

- Ha veszünk 2 kg almát, 1 kg epret és 1,5 kg meggyet:

[2, 0, 0, 1, 1.5, 0]

Gyümölcs	kód
alma	0
körte	1
barack	2
eper	3
meggy	4
kivi	5

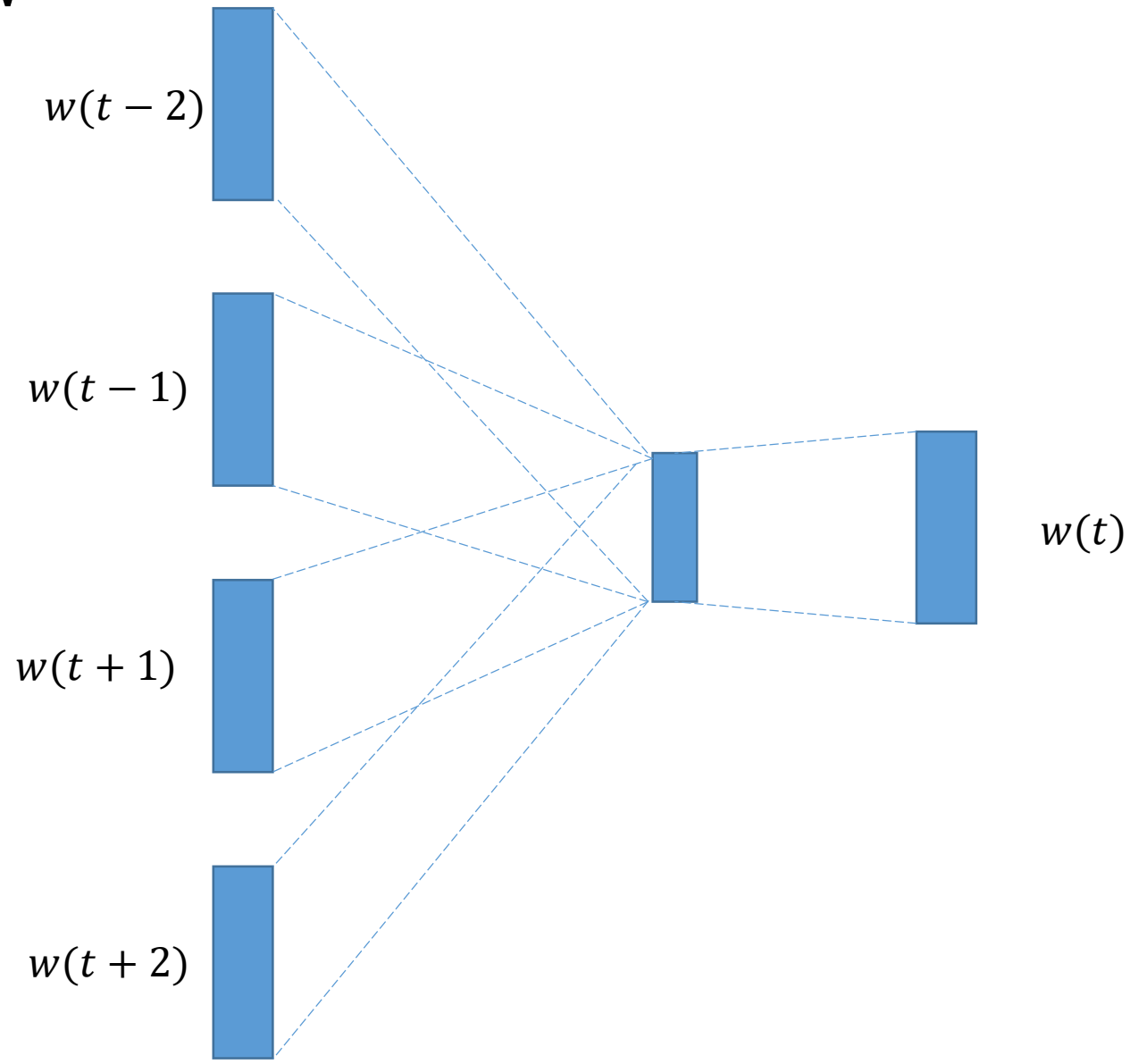
Szózsák (bag of words)

- Lényegében egy leltár: halmaz, mely az egyes elemek számosságát is tartalmazza

- Pl. automatikus tartalmi kivonatolásban szövegre:
- egy vektor
- annyi dimenziós, ahány szó van (a szótárunkban)
- szavakat egy (egész) számra képezzük (0..N, ha N szó van)
- minden pozícióban az adott szó előfordulásának száma a szövegben
- -> vektorokkal leír egy egységet
- -> sorrendiségi információk elvesznek (kontextus marad)

- Vegyük észre, hogy ez egyben kódolás is.

Cbow



Távolságmetrikák

- Koszinusz-hasonlóság

$$cos\varphi = \frac{A\ B}{||A||||B||} = \frac{\sum_k a_k b_k}{\sqrt{\sum_k {a_k}^2}\ \sqrt{\sum_k {b_k}^2}}$$

- Jaccard-hasonlóság

$$J(A,B) = \frac{|A \cap B\ |}{|A \cup B\ |}$$

Szózsák

- 1. dokumentum:
  Az idő szép ma is. Holnap nem fog esni.
- 2. dokumentum:
  Szép idő volt ma. Holnap is szép lesz az idő.

	az	idő	szép	ma	is	holnap	nem	fog	esni	volt	lesz
1.dok.	1	1	1	1	1	1	1	1	1	0	0
2.dok.	1	2	2	1	1	1	0	0	0	1	1

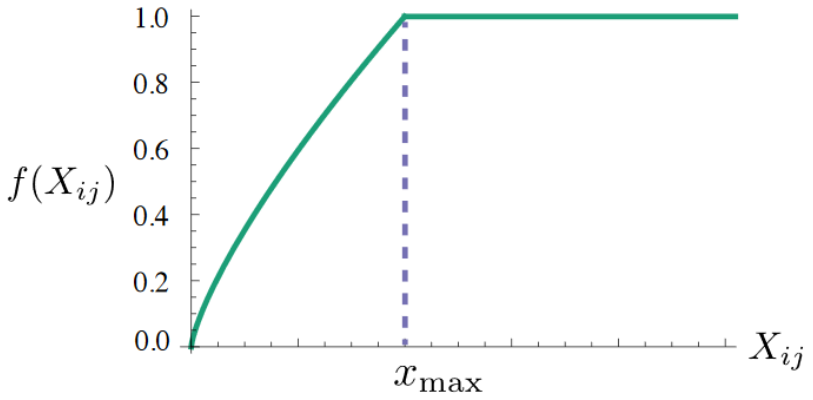
Szavak távolsága

- http://bionlp-www.utu.fi/wv_demo/
- https://lamyiowce.github.io/word2viz/

GloVe

- https://nlp.stanford.edu/pubs/glove.pdf
- Módosított költségfüggvény:

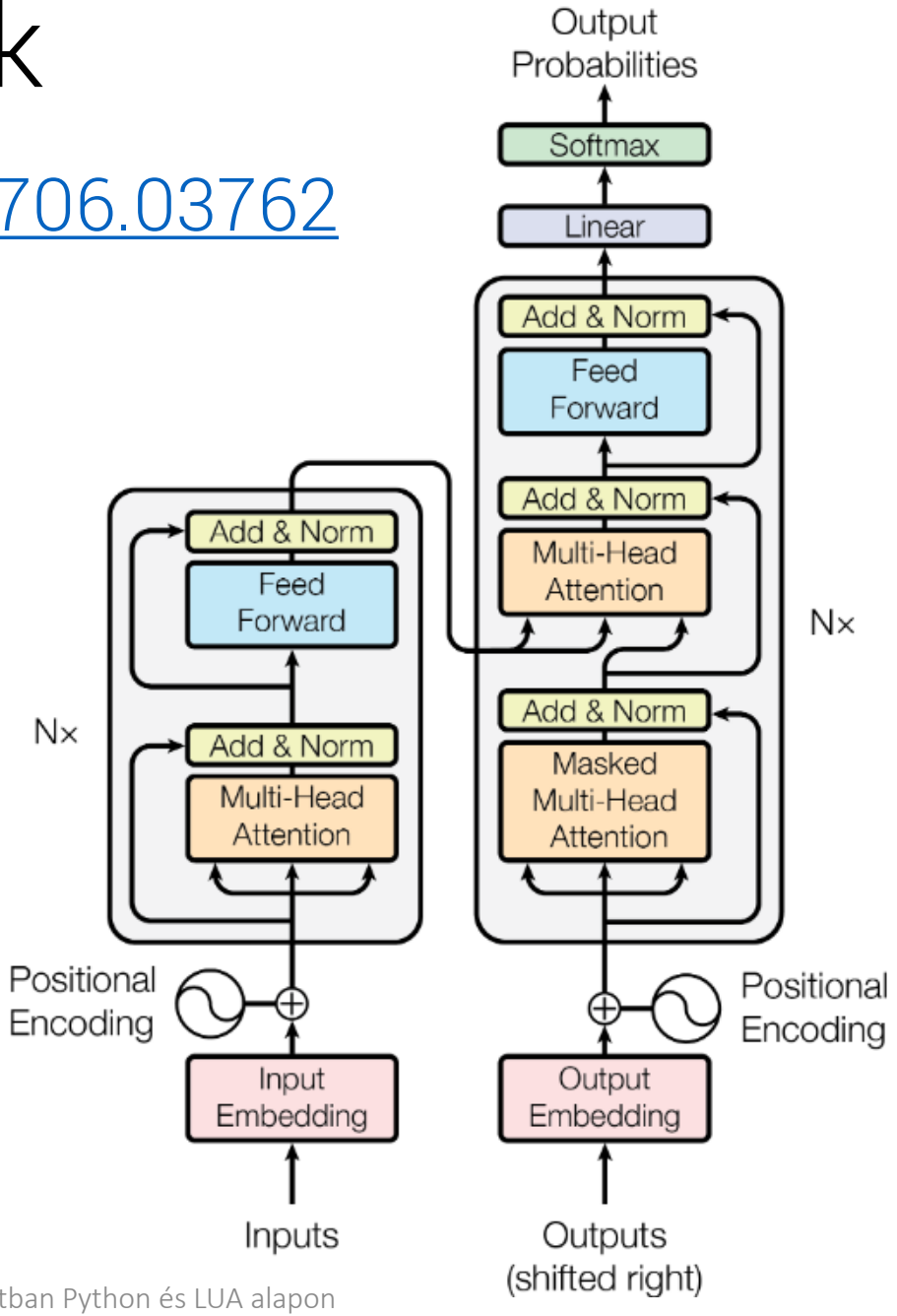
$$J(\theta) = \frac{1}{2} \sum_{i,j=1}^{W} f(P_{ij})(u_i^T v_j - \log P_{ij})^2$$



Seq-to-seq

Transzformerek

- https://arxiv.org/abs/1706.03762



BERT

- Bidirectional Encoder Representations from Transformers
  - Pre-trained Language Model
- ELMO
- OpenAI GPT

- MLM (Masked Language Models)
- NSP (Next Sentence Prediction)

BERT

- 2 lépéses tanítás
  - Pre-Training (unsupervise)
    - NSP
    - Mask LM
    - Big dataset
  - Fine-Tuning
    - Small dataset

BERT

- Input (token, segment, position embedding)
- Pre-train
  - MLM
    - 15% MASK
      - 80% MASK
      - 10% random token
      - 10% original token
  - NSP (next sentence)
    - Binary classification
- Fine-tuning
  - Single Sentence Classification (target label) pl: sent,+/-
  - Single Sentenve Tagging (pl.slot filling time, dest, other)
  - Sentence Pair Classification
  - Question Answering Task

BERT

- https://huggingface.co/bert-base-uncased

GPT-3

- https://arxiv.org/pdf/2005.14165.pdf
- OpenAI (GPT-n)
- 175 milliárd paraméter

- https://medium.com/towards-artificial-intelligence/crazy-gpt-3-use-cases-232c22142044

Kiértékelés

- Önállóan
  - Metrikák
  - Nem garatált, hogy a használatban jó lesz
- Beépítve, beágyazva
  - Összetett hatás
  - Beágyazás volt-e a javító tényező

Forrás: Jeffrey Pennington, Richard Socher, Christopher D. Mann: GloVe: Global Vectors for Word Representatio https://nlp.stanford.edu/pubs/glove.pdf

Kiértékelés

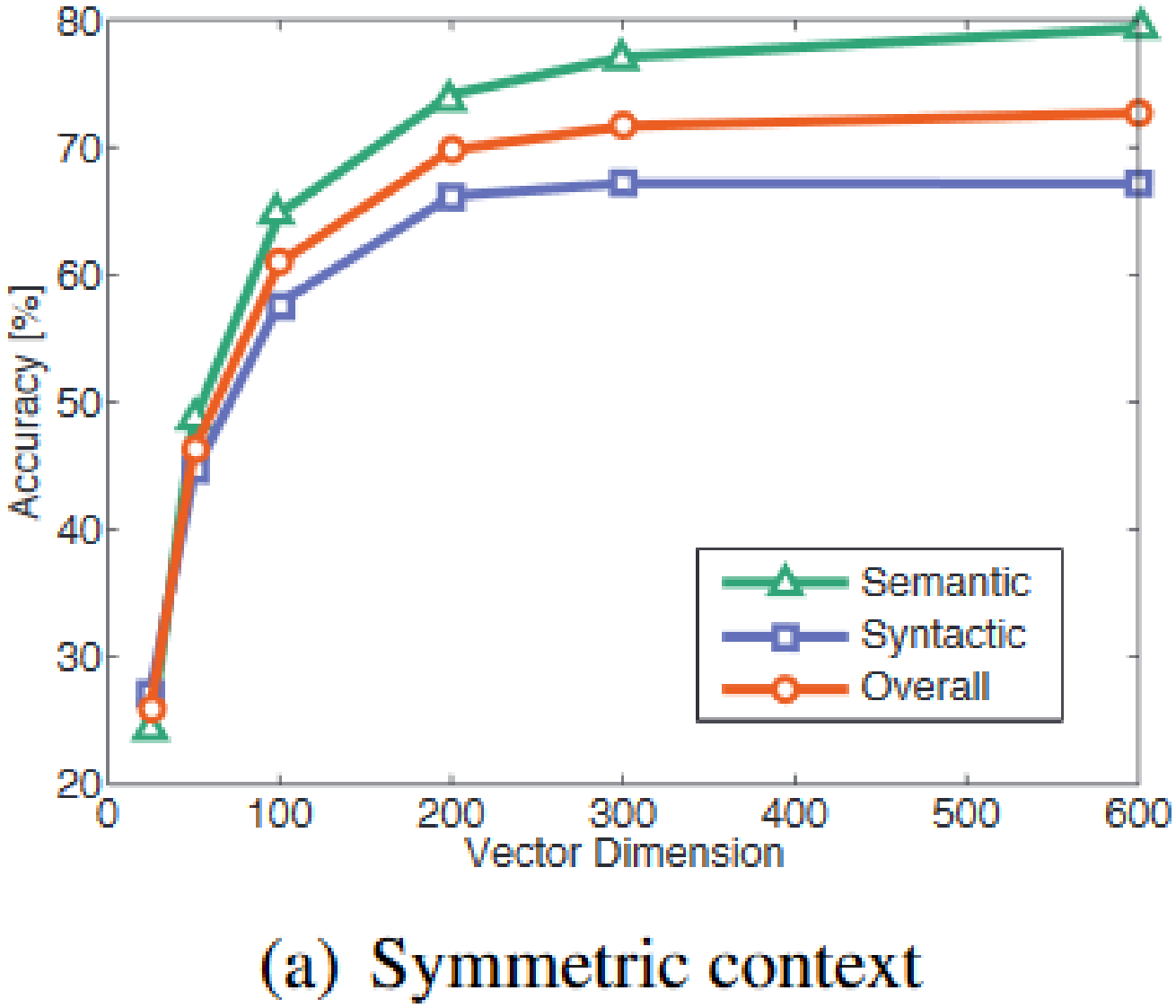
athens greece berlin germany
moscow russia sofia bulgaria
pittsburgh pennsylvania orlando florida
croatia kuna poland zloty
brothers sisters prince princess
complete completely most mostly
competitive uncompetitive known unknown
bad worse low lower
big biggest long longest
debug debugging fly flying
australia australian switzerland swiss
decreasing decreased sitting sat
banana bananas dog dogs
eat eats say says

GloVe kiértékelése

Model	Dim.	Size	Sem.	Syn.	Tot.
ivLBL	100	1.5B	55.9	50.1	53.2
HPCA	100	1.6B	4.2	16.4	10.8
GloVe	100	1.6B	67.5	54.3	60.3
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
vLBL	300	1.5B	54.2	64.8	60.0
ivLBL	300	1.5B	65.2	63.0	64.0
GloVe	300	1.6B	80.8	61.5	70.3
SVD	300	6B	6.3	8.1	7.3
SVD-S	300	6B	36.7	46.6	42.1
SVD-L	300	6B	56.6	63.0	60.1
CBOW†	300	6B	63.6	67.4	65.7
SG†	300	6B	73.0	66.0	69.1
GloVe	300	6B	77.4	67.0	71.7
CBOW	1000	6B	57.3	68.9	63.7
SG	1000	6B	66.1	65.1	65.6
SVD-L	300	42B	38.4	58.2	49.2
GloVe	300	42B	**81.9**	**69.3**	**75.0**

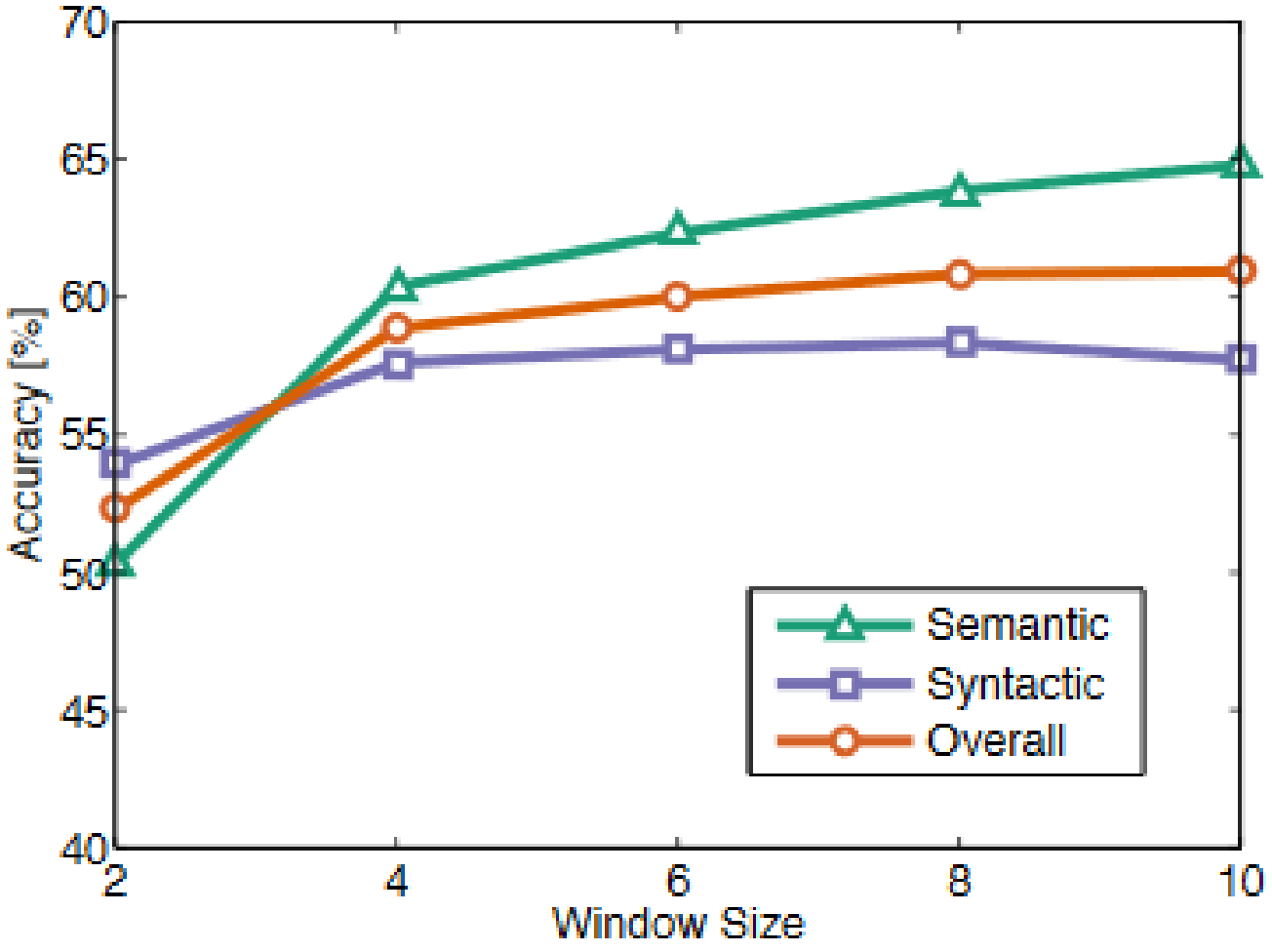
- Forrás: Jeffrey Pennington, Richard Socher, Christopher D. Mann: GloVe: Global Vectors for Word Representatio https://nlp.stanford.edu/pubs/glove.pdf

GloVe



- Forrás: Jeffrey Pennington, Richard Socher, Christopher D. Mann: GloVe: Global Vectors for Word Representatio https://nlp.stanford.edu/pubs/glove.pdf

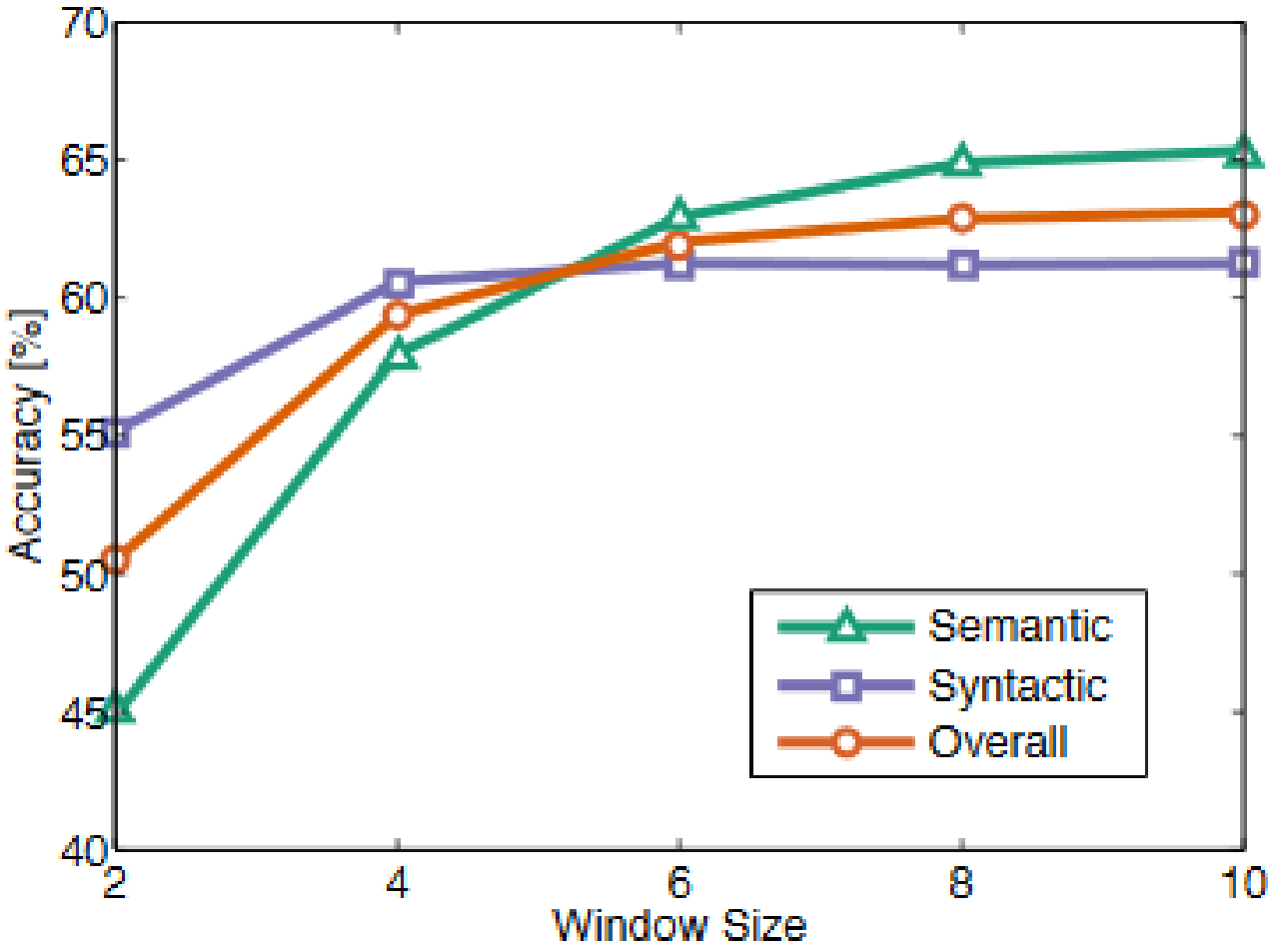
GloVe



(b) Symmetric context

- Forrás: Jeffrey Pennington, Richard Socher, Christopher D. Mann: GloVe: Global Vectors for Word Representatio https://nlp.stanford.edu/pubs/glove.pdf

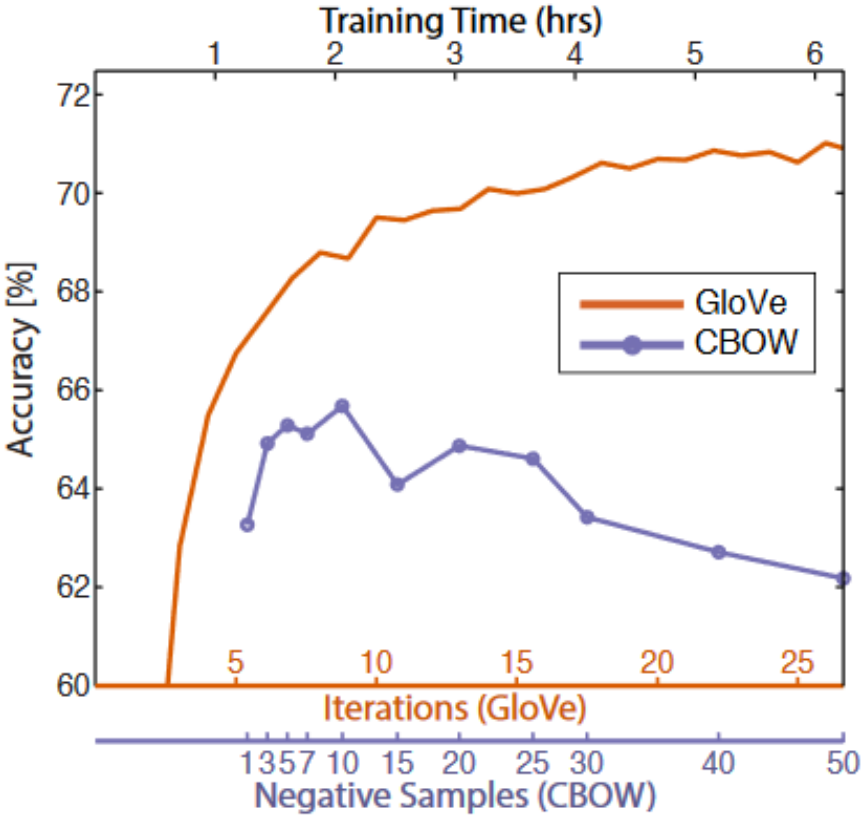
GloVe



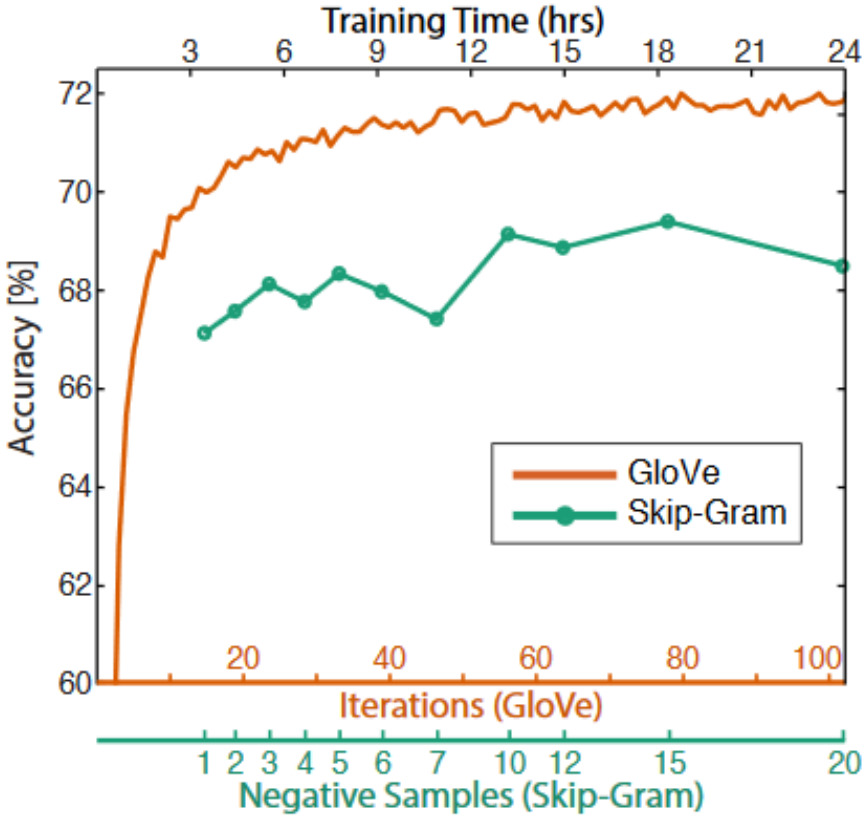
(c) Asymmetric context

- Forrás: Jeffrey Pennington, Richard Socher, Christopher D. Mann: GloVe: Global Vectors for Word Representatio https://nlp.stanford.edu/pubs/glove.pdf

GloVe



(a) GloVe vs CBOW



(b) GloVe vs Skip-Gram

- Forrás: Jeffrey Pennington, Richard Socher, Christopher D. Mann: GloVe: Global Vectors for Word Representatio https://nlp.stanford.edu/pubs/glove.pdf

Beágyazás

- Bottleneck (autoencoder)
- Skip-gram, CBOW, N-gram, GloVe

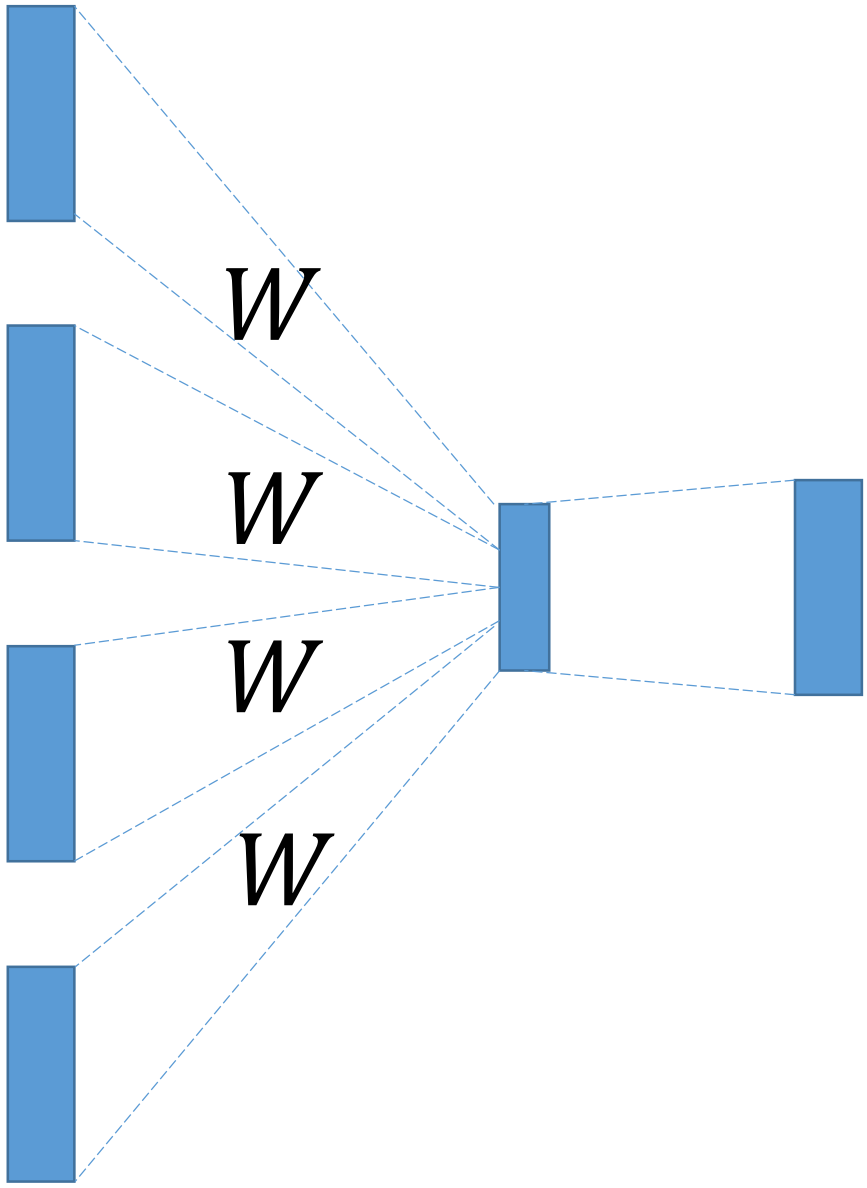
- Tanításkor a hálót optimalizáljuk valamilyen feladatra
  - Szűkített réteget használjuk fel
  - Tömörítés
  - Dimenzió csökkentés
  - Transfer learning



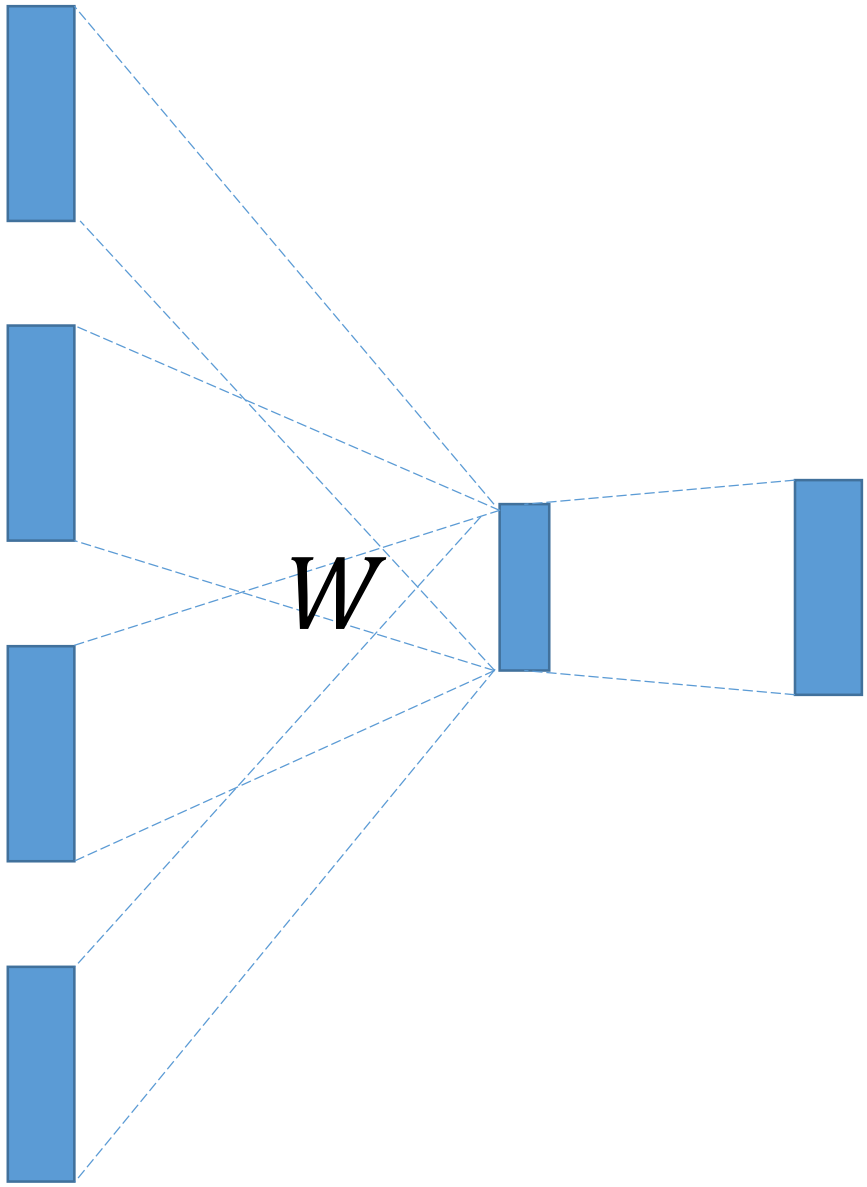
Névnek a NEPTUN kódodat add meg!

https://kahoot.it/

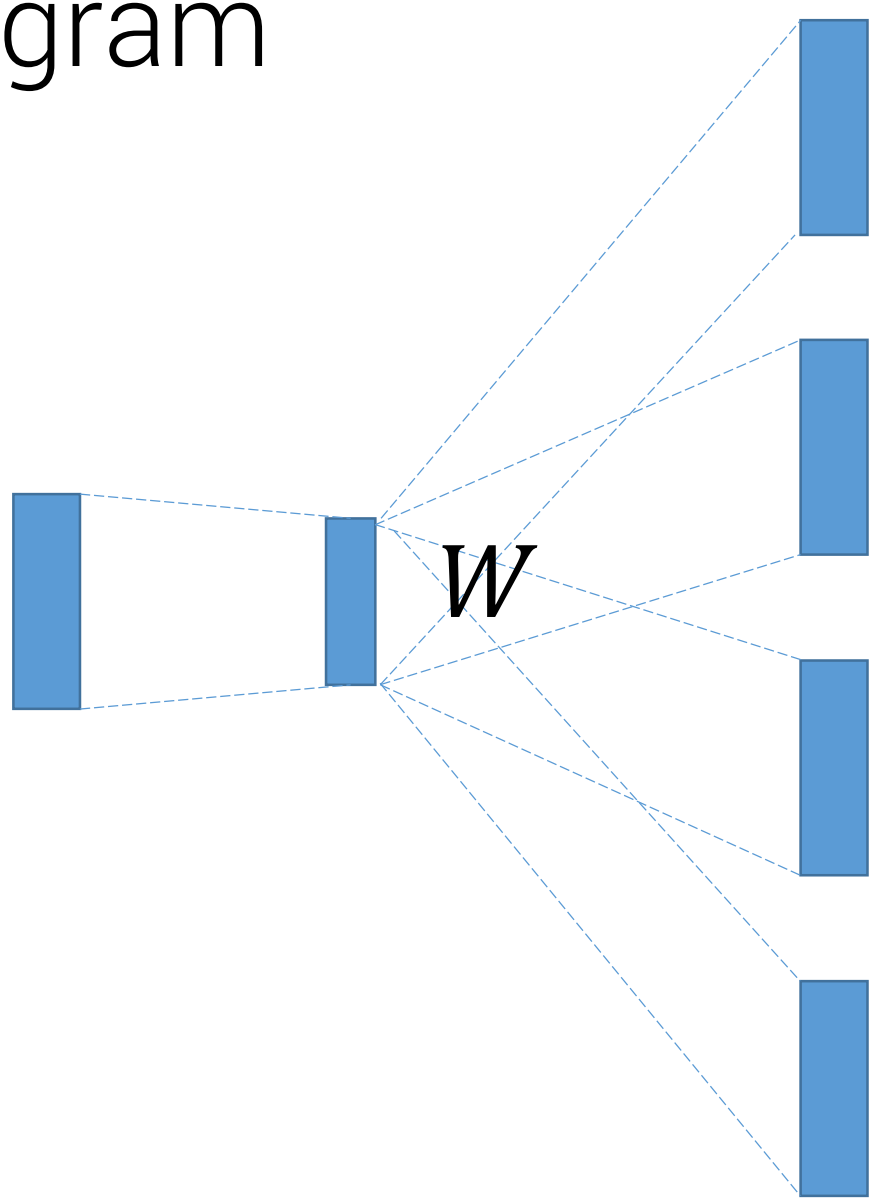
N-gram



Cbow



Skip-gram



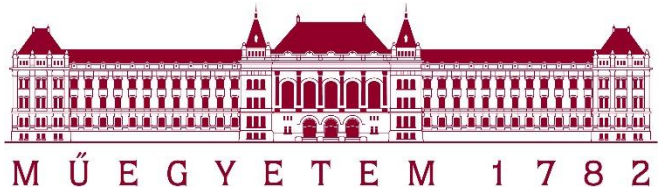
t-SNE

t-Distributed Stochastic Neighbor Embedding



Forrás: https://towardsdatascience.com/t-sne-python-example-1ded9953f26







Köszönöm a figyelmet!

zainko@tmit.bme.hu



http://smartlab.tmit.bme.hu

