Jezični model temeljen na neuronskoj mreži

Florijan Stamenković
Mentor: Marko Čupić

Fakultet elektrotehnike i računarstva
Sveučilište u Zagrebu

July 13, 2015

Jezični model

Jezični model

Smislene rečenice

Petar ide u dućan. Danas je lijep dan.

Jezični model

Smislene rečenice

Petar ide u dućan. Danas je lijep dan.

Besmislene rečenice

Gavran tanges jučer zelen dva.

Jezični model

Smislene rečenice

Petar ide u dućan. Danas je lijep dan.

Besmislene rečenice

Gavran tanges jučer zelen dva.

Matematička (probabilistička) definicija

$P(polje|Petar, ide, u) > P(vilica|Petar, ide, u)$

$P(w_i|w_0, w_1, ..., w_{i-1})$

Jezični model

Smislene rečenice

Petar ide u dućan. Danas je lijep dan.

Besmislene rečenice

Gavran tanges jučer zelen dva.

Matematička (probabilistička) definicija

$P(polje|Petar, ide, u) > P(vilica|Petar, ide, u)$
$P(w_i|w_0, w_1, ..., w_{i-1})$

Aproksimacija korištenjem $(n-1)$ prethodnih riječi

$P(w_i|w_{i-n+1}, ..., w_{i-1}) \approx P(w_i|w_0, w_1, ..., w_{i-1})$

Jezični model

Primjene

- Automatsko prevođenje

Jezični model

Primjene

- Automatsko prevođenje
- Prepoznavanje govora

Jezični model

Primjene

- Automatsko prevođenje
- Prepoznavanje govora
- Ispravka pogreški u pisanju

Jezični model

Primjene

- Automatsko prevođenje
- Prepoznavanje govora
- Ispravka pogreški u pisanju
- Klasifikacija teksta

Jezični model

Primjene

- Automatsko prevođenje
- Prepoznavanje govora
- Ispravka pogreški u pisanju
- Klasifikacija teksta
- ...

Prebrojavanje n-grama

Primjerice trigrama

- Primjer: Škola je uskoro gotova.
- Trigrami: (škola je), (je uskoro), (uskoro gotova)
- $P(w_i|w_{i-2}, w_{i-1}) \propto C(w_{i-2}, w_{i-1}, w_i)$

Prednosti i nedostatci

- Jednostavan pristup
- Brz za implementaciju, treniranje i primjenu
- Rijetkost pojavljivanja onemogućava korištenje velikih n
- Potrebno dobro (Kneser-Ney) zaglađivanje da bi radilo

Neuronska mreža

Definicija

- Unaprijedna mreža bez skrivenih slojeva
- Po jedan izlazni neuron za svaku riječ vokabulara
- Običan klasifikator, riječ je klasa
- Riječi predočavane d-dimenzionalnim vektorima

Prednosti i nedostatci

- Veliki parametarski prostor (veličina vokabulara)
- Sporo treniranje, osrednji rezultati

Log-bilinearni model

Definicija

- Riječi predočene d-dimenzionalnim vektorima
- Prethodnih $(n-1)$ riječ daju $(n-1)d$ vektor
- Dobiveni vektor se množi matricom W, izlaz je d-dimenzionalni vektor
- Sličnost izlaznog vektora definira vjerojatnost riječi

Prednosti i nedostatci

- Najbolji rezultati, riječ-vektori su nusprodukt
- Jednostavna formulacija
- Rezultati samo malo bolji od Kneser-Ney, puno sporije treniranje

Ograničeni Boltzmannov stroj

Definicija

- Riječi predočene d-dimenzionalnim vektorima
- Jedan skriveni sloj stohastičkih binarnih neurona
- Energija mreže definira uvjetnu vjerojatnost

Prednosti i nedostatci

- Kompleksno i sporo treniranje stohastičkih neurona
- Osrednji rezultati

Rezultati

Evaluacija mjerom perplexity

- $\exp\left(\frac{1}{N} \ln P(w|...)\right)$
- Manje je bolje
- Jako ovisi o vokabularu, korpusu; samo za relativnu usporedbu

	Parametar n		
Model	3	4	5
Additivno	305	1182	2680
Kneser-Ney	72	121	204
Neuronska mreža	117	114	113
Log-biliner	102	98	98

Rezultati

Trajanje treniranja

- CPU: Intel i7, 3.5GHz, 4-core, 8-thread
- GPU: Nvidia GTX 960M, 4GB, 640-cores

Table: Trajanje treniranja modela, izraženo u SAT:MINUTE obliku.

	Parametar n		
Model	3	4	5
Prebrojavanje n-grama CPU	0:01	0:01	0.01
Neuronska mreža, GPU	1:47	2:15	2:17
Neuronska mreža, CPU	13:38	13:52	14.20
Log-bilinear GPU	0:49	0:48	0:46
Log-bilinear CPU	8:07	8:01	6:56

Demo