

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

$$P(\textit{polje} | \textit{Petar}, \textit{ide}, \textit{u}) > P(\textit{vilica} | \textit{Petar}, \textit{ide}, \textit{u})$$

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

Jezični model

Smislene rečenice

Petar ide u dućan. Danas je lijep dan.

Besmislene rečenice

Gavran tanges jučer zelen dva.

Matematička (probabilistička) definicija

$$P(\textit{polje} | \textit{Petar}, \textit{ide}, u) > P(\textit{vilica} | \textit{Petar}, \textit{ide}, u)$$

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

Aproksimacija korištenjem $(n - 1)$ prethodnih riječi

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

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 trigramama

- ▶ 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

Demo