Automatsko prevođenje

Seminarski rad u okviru kursa Računarska Inteligencija

Matematički fakultet, Univerzitet u Beogradu

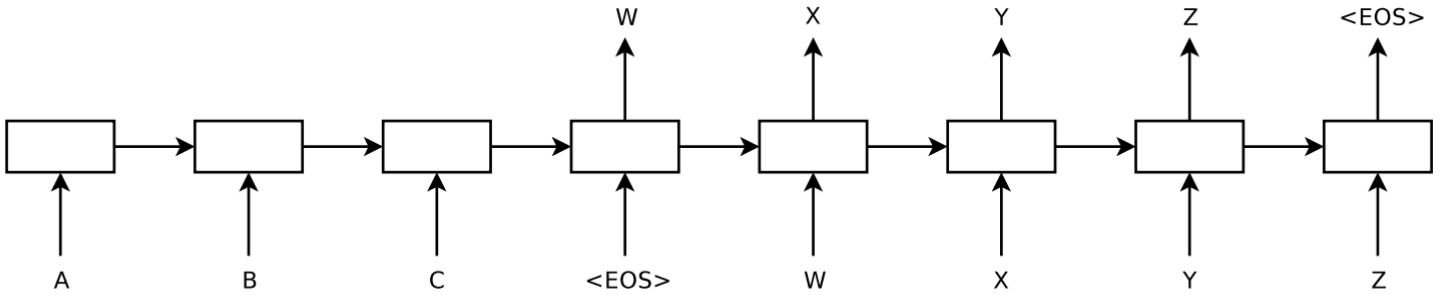
Vladimir Vuksanović 145/2017

Šta je automatsko prevođenje?

Automatsko prevođenje podrazumeva prevođenje teksta sa jednog prirodnog jezika na drugi korišćenjem kompjutera. Preciznije, potrebno je za ulaznu rečenicu proizvoljne dužine na izvornom jeziku generisati izlaznu rečenicu, ne nužno iste dužine, na ciljnom jeziku koja ima isto značenje kao ulazna.

Rešenje

- **Koristićemo neuronsku mrežu sa enkoder-dekoder arhitekturom i LSTM slojevima**



- **Ideja: Mreža se podeli na dva dela: enkoder i dekoder**
- **Enkoder od rečenice na izvornom jeziku generiše reprezentaciju fiksne dužine (misao)**
- **Dekoder od te reprezentacije iterativno konstruiše rečenicu na ciljnom jeziku reč po reč**

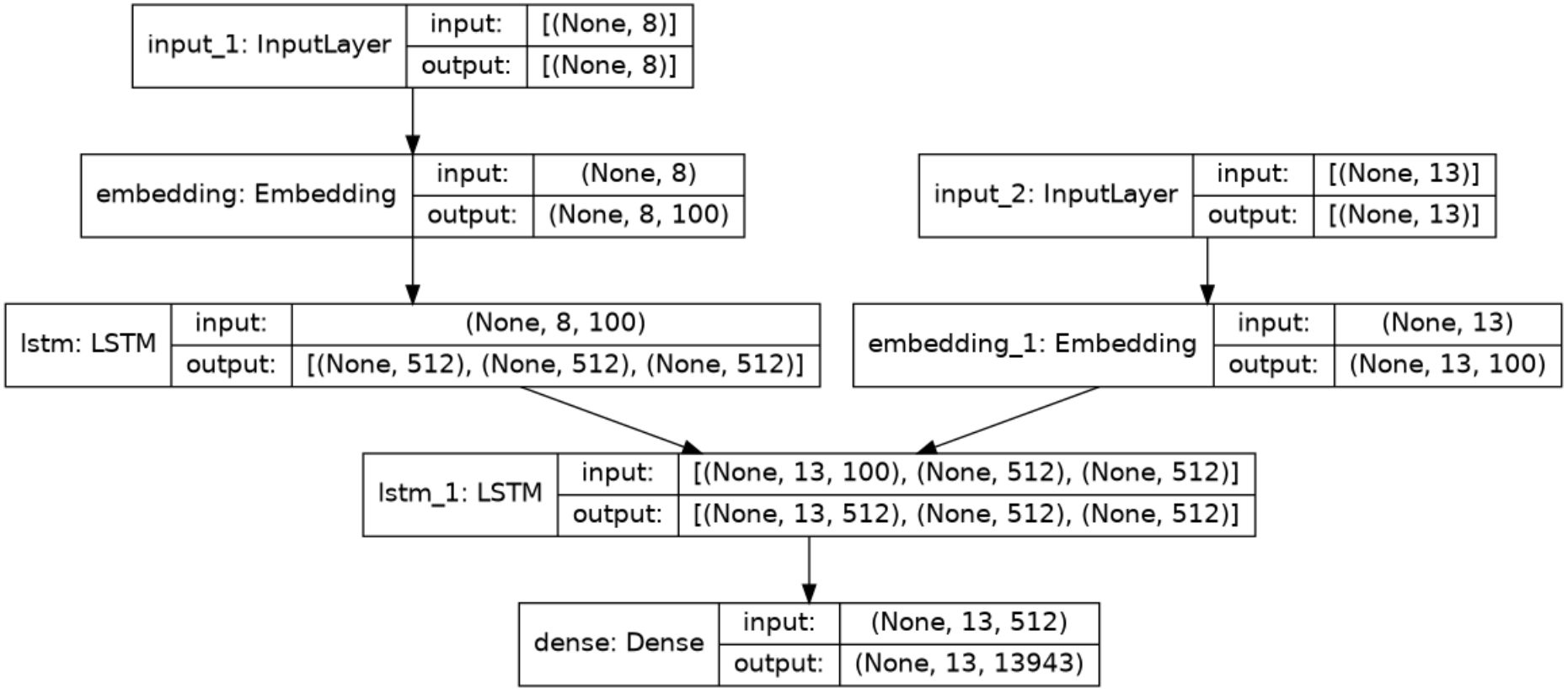
Pretprocesiranje

Skup podataka čine parovi rečenica na engleskom i francuskom jeziku. Da bi se koristile za treniranje mreže potrebno je pretvoriti ih u numeričke nizove. Postupak je sledeći:

1) **Eliminiacija duplikata**
2) **Izbacivanje specijalnih simbola**
3) **Tokenizacija**
4) **Obrtanje redosleda tokena ulaznih rečenica (opciono)**
5) **Dopunjavanje izlaznih rečenica terminirajućim tokenima**
6) **Dopunjavanje nizova do iste dužine**
7) **Konvertovanje izlaznih reči u kategorijsku promenljivu**

I'm pleased to see you. → [0 0 0 12 1144 7 81 2]

Model za treniranje (1)

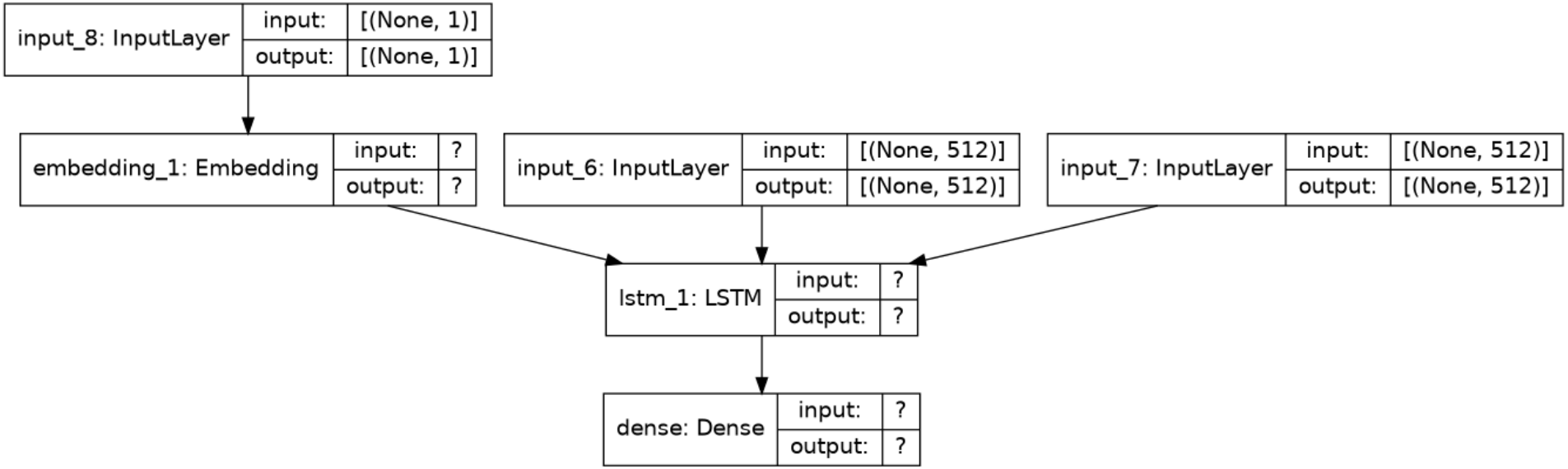


Model za treniranje (2)

- 50.000 rečenica
- Embedding sloj dimenzije 100
- GloVe težine za embedding na engleskom jeziku
- LSTM sloj dimenzije 512
- Optimizator: rmsprop
- Funkcija greške: unakrsna kategorijska entropija
- Metrika: tačnost
- 5 epoha
- Batch size: 128
- Validacija na 10% trening skupa

Model za predviđanje

Prethodni model zahteva da je ceo prevod poznat od samog početka. To je korisno prilikom treniranja ali neupotrebljivo za prevođenje novih rečenica. Zbog toga modifikujemo dekoder da prima po jednu reč i pozivamo ga iterativno do kraja prevoda.



Beam search algoritam

- **Heuristika koja paralelno razmatra više prevoda i bira najbolji među njima**
- **Maksimizira se formula** $\mathrm{argmax}_y \frac{1}{T_y} \sum_{t=1}^{T_y} \log P\left(y^{\langle t \rangle} \middle| X, y^{\langle 1 \rangle}, y^{\langle 2 \rangle}, \ldots, y^{\langle t-1 \rangle}\right)$
- **Bira se k reči koje mreža predviđa kao najverovatnije na osnovu vrednosti potpuno povezanog sloja.**
- **Kada se izabere prvih k kandidata, svaki od njih se potom dopunjuje svim mogućim rečima i bira se k najverovatnijih sevenci među njima.**
- **Ovaj postupak se ponavlja do kraja svake od sekvenci i onda se za konačan prevod uzima najbolja od njih.**
- **Specijalno, za vrednost k = 1 postupak se svodi na gramzivu pretragu.**

Rezultati na trening podacima (1)

Rečenica	What's that dog doing?
Tačan prevod	Que fait ce chien ?
Model (greedy)	de quoi chien est il
Model (beam)	qu'est ce que ce chien

Rečenica	Tom tasted the cake.
Tačan prevod	Tom a goûté le gâteau.
Model (greedy)	tom a mangé le gâteau
Model (beam)	tom a mangé le gâteau

Rezultati na trening podacima (2)

Rečenica	You can't bury the truth.
Tačan prevod	Vous ne pouvez pas enterrer la vérité.
Model (greedy)	vous ne pouvez pas les autres
Model (beam)	vous ne pouvez pas la vérité

Rečenica	That boy is smart.
Tačan prevod	Ce garçon est intelligent.
Model (greedy)	ce garçon est intelligent
Model (beam)	ce garçon est intelligent

Rezultati na proizvoljnim rečenicama

Rečenica	What time is it?
Tačan prevod	Quelle heure est il?
Model (greedy)	à quoi ça est le temps
Model (beam)	à quelle heure est il

Rečenica	I am going to run home tomorrow.
Tačan prevod	Je vais courir à la maison demain.
Model (greedy)	je vais aller à la maison
Model (beam)	je vais aller à la maison

BLEU ocena

BLEU (Bilingual Evaluation Understudy) ocena je mera kvaliteta mašinskog prevoda. Ima vrednost između 0 i 1.

Rečenica	You're polite.	
Tačan prevod	tu es poli	
Model	tu es intelligent	1.13e^-154

Rečenica	That sounds exhausting.	
Tačan prevod	ça a l'air épuisant	
Model	ça a l'air bizarre	8.63e^-78

BLEU ocena

Rečenica	I'm as tall as my father.	
Tačan prevod	je suis aussi grand que mon père	
Model	je suis aussi grand que moi	0.64

Rečenica	I didn't sleep.	
Tačan prevod	je n'ai pas dormi	
Model	je n'ai pas dormi	1.0

Dalje unapređivanje

- **Veći skup podataka za trening (WMT14)**
- **Više LSTM slojeva**
- **Obrtanje redosleda reči na ulaznim rečenicama**
- **Mehanizam za poboljšanje performansi na retkim rečima**

Hvala na pažnji!