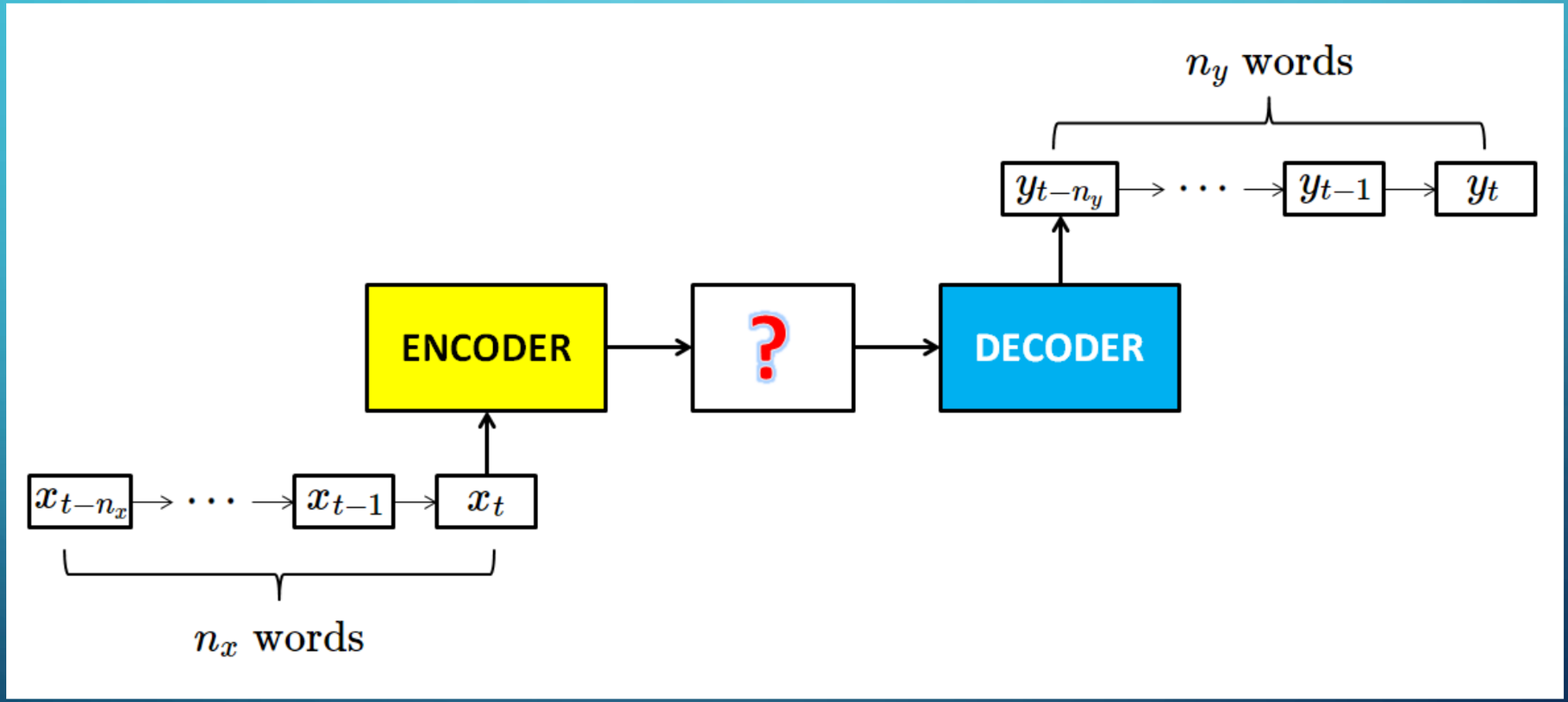
Neuronske mreže 2017

Sequence To Sequence

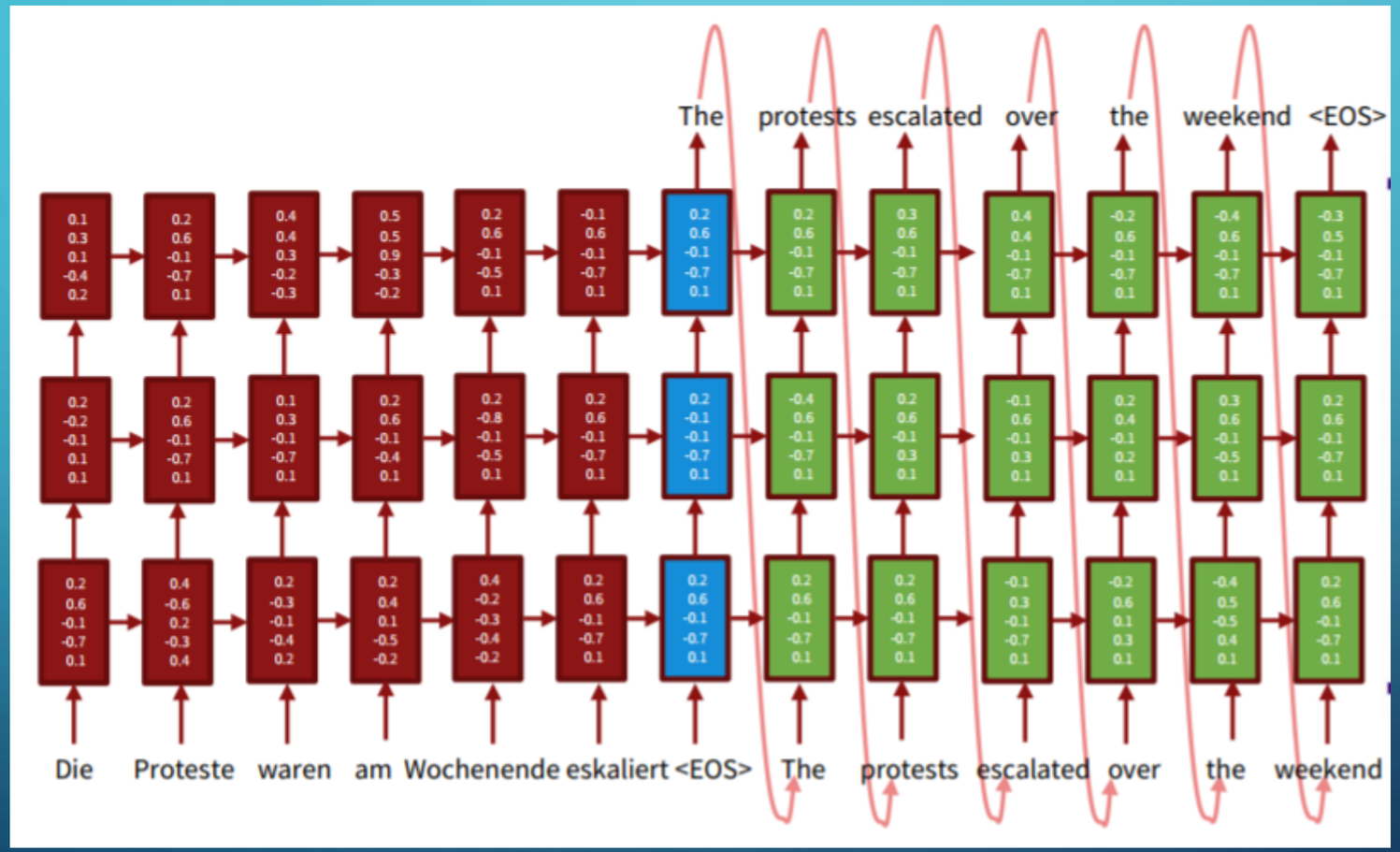
Sequence To Sequence

- Model predstavljen u radu https://arxiv.org/pdf/1409.3215.pdf
- **Ideja**: na osnovu ulazne sekvence generisati izlaznu sekvencu
- **Enkoder** – **Dekoder** arhitektura
- Primena: *chat bot, machine translation, text summarization, question answering, speech to text..*

Enkoder – Dekoder arhitektura



Enkoder – Dekoder arhitektura

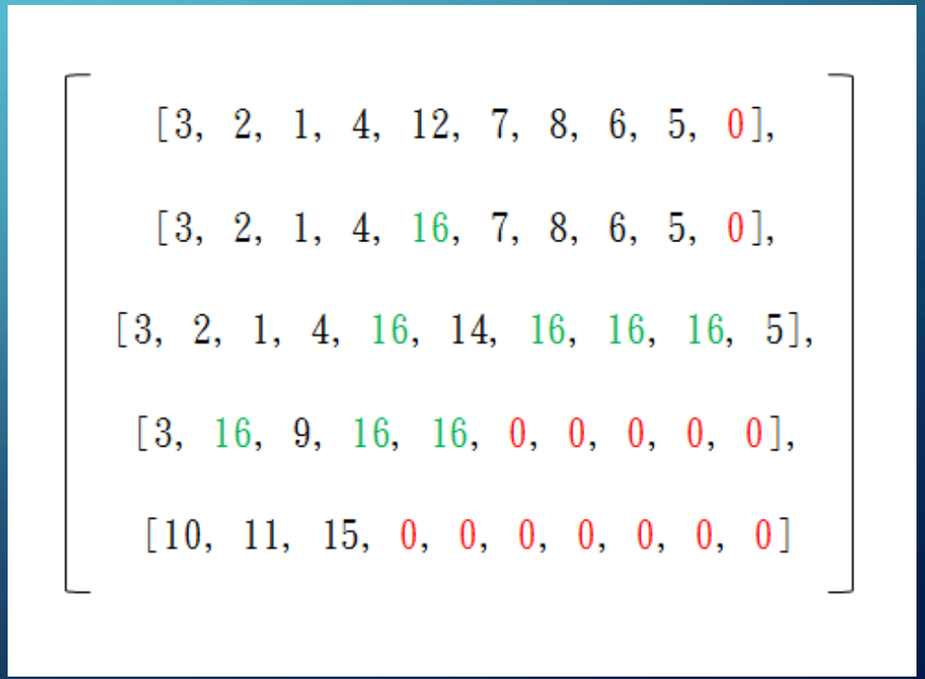


Enkoder – Dekoder arhitektura

- Enkoder – predstavljen najčešće preko LSTM mreze
  - Ulaz: sekvenca koju želimo da enkodujemo
  - Izlaz: enkodovanja stanje
- Dekoder – LSTM
  - Ulaz: enkodovano stanje
  - Izlaz: dekodirana sekvenca

Enkoder – Dekoder arhitektura

- Kako transofrmisati tekst u vektor fiksne dužine?
  - ZERO PADDING – dopuniti kraće sekvence
  - Nepoznate reči
  - Kreiranja rečnika
  - Svaka reč opisan svojim indeksom u rečniku



Proširenje

- *Word2Vec i GloVe (Embeddings)*
- *Reverse input*
- *Bidirectional LSTM*
- *Attention mechanisim*
- *Teacher forcing*
- *Pointer networks*
- *Beam search*

Reprezentacija reči

- *Word2Vec i GloVe – reči pozicionirane u N dimenzionalnom prostoru*
- *Reprezentacija pomoću rečnika – nemamo informaciju u kakvom su odnosu reči*
- *Nema semantika*
- *Word2Vec i GloVe opisuje reči N dimenzionalnim vektorima*
- King - Man + Woman = Queen
- *Demo link*

Reverse input

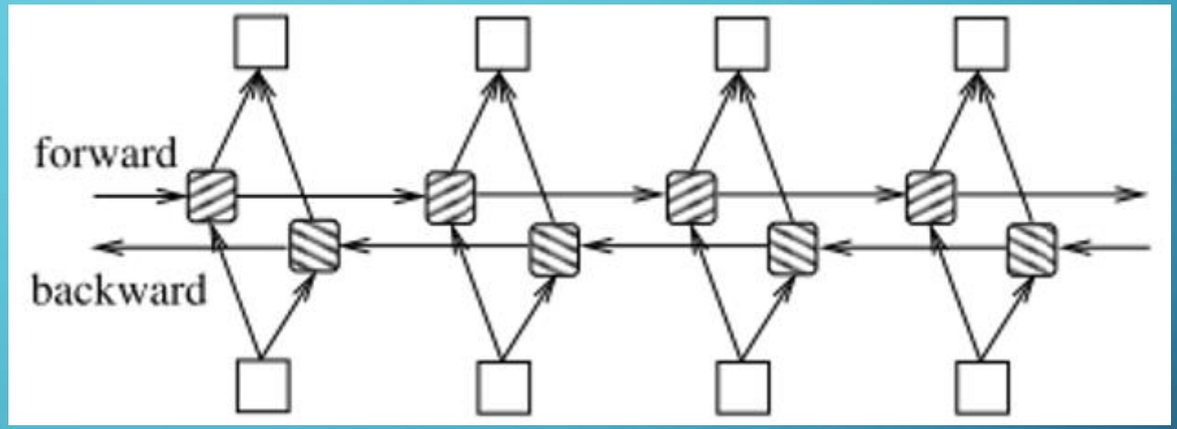
- Ulazni vektor u obrnutom redosledu daje mnogo bolje rezultate

- A,B,C → a, b, c mreža će mnogo brže naučiti ako zapišemo: C, B, A → a, b, c
- SGD mnogo brže konvergira

Bidirectional LSTM

- 2 LSTM mreže, za svaki smer po jedna

Višeslojna BiLSTM arhitektura (Stacked)
Zahtevno za obučavanje

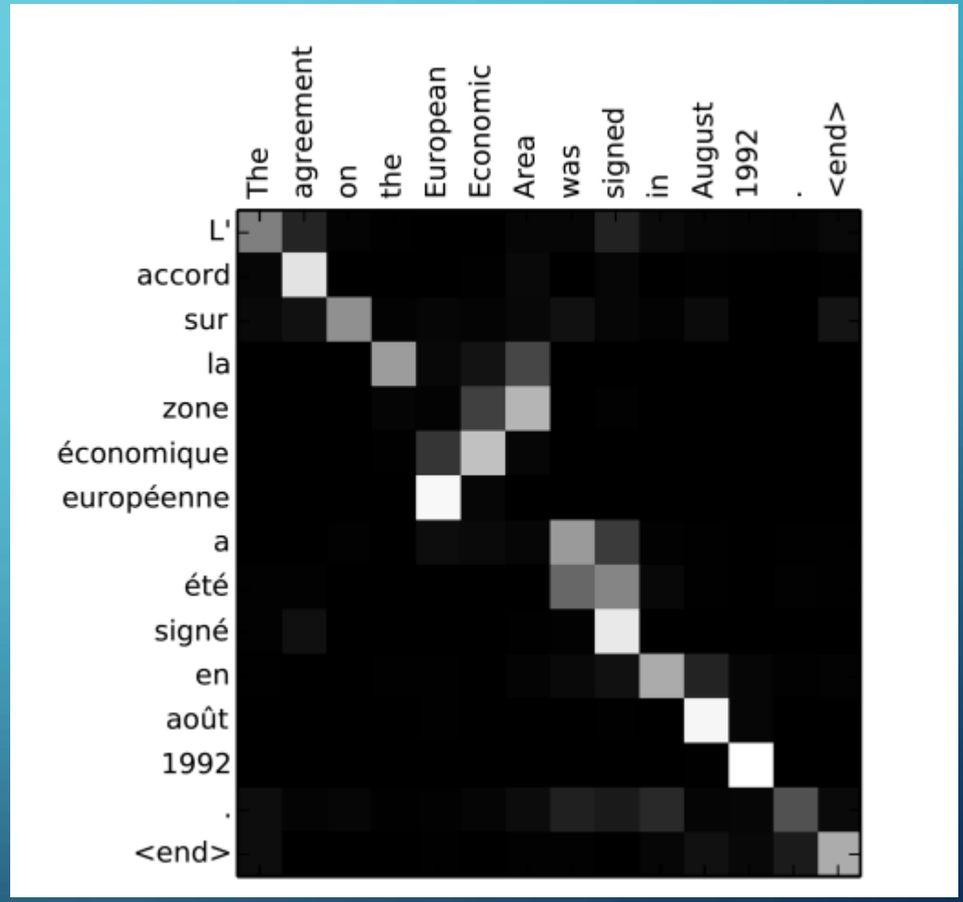


Attention mechanism

- LSTM ne može da zapamti sekvence duže od 30 karaktera
- Dozvoliti dekoder delu da “bira” delove sekvence iz enkodera

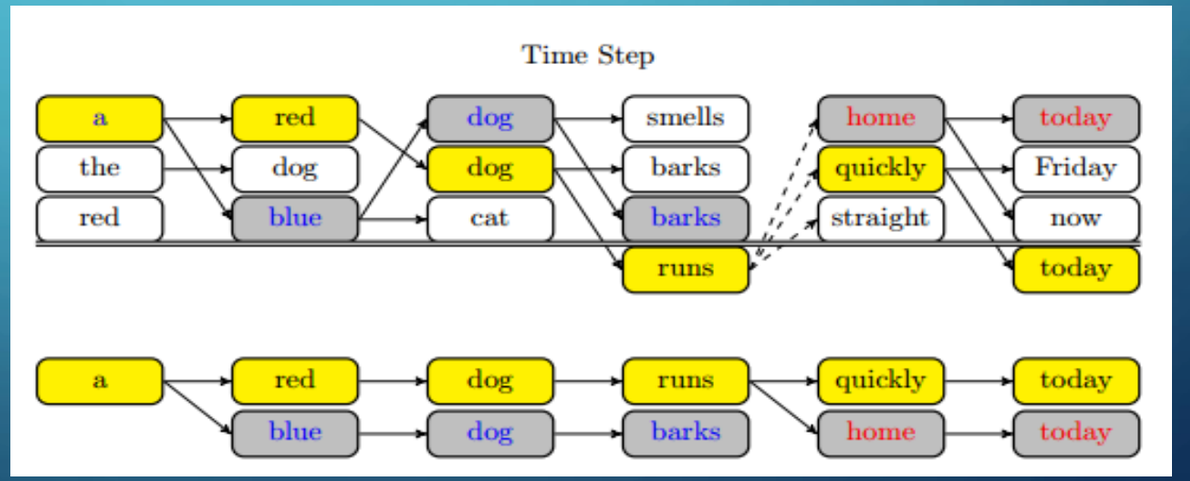
- Kao RAM fiksne veličine
- Lokalni vs Globalni

Detaljnije na linku



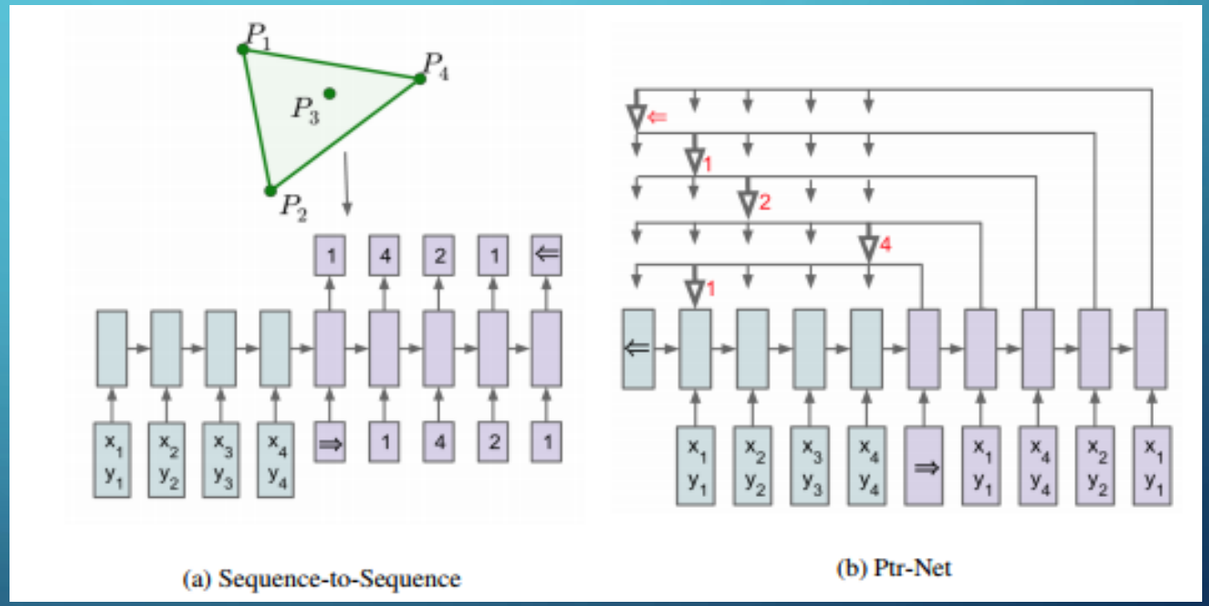
Beam search

- Upotreba Greedy pretrage pri dekodiranju može uticati na tačnost modela
  - Greedy uvek pokupi prvi najbolji rezultat, tok ostale odbaci
- Upotrebom Beam pretrage se uzima K najboljih rezultata, a nakon toga se pretragom određuje najbolje rešenje
  - Računski kompleksno, ali daje bolje rezultate u odnosu na klasičnu pretragu
  - Standard u NMT



Pointer networks

- Predstavljaju nadogradnju Sequence to Sequence modela
- Za razliku od seq2seq, izlaz nije dekodirana sekvenca nego index(i) ulazne sekvence
- Primena u Question Answering modelima
- Više o samoj ideji u radu na linku



Metrike

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
- BLEU (Bilingual evaluation understudy**)**

Datasets

- http://www.manythings.org/anki/
- https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html
- http://cs.nyu.edu/~kcho/DMQA/
- http://www.statmt.org/europarl/
- Gigaword i DUC (nisu free)

Primeri

- Translation ENG – SR
- Text summarization
- Chat bot

Dodatni materijali

- http://web.stanford.edu/class/cs224n/lectures/cs224n-2017-lecture10.pdf
- https://github.com/tensorflow/nmt
- http://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html
- https://github.com/samwit/TensorFlowTalks/blob/master/Talk05_Seq2Seq_NMT/TensorFlow05-Neural-translation.pdf