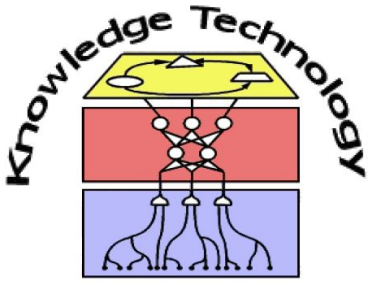
Sequence-to-Sequence Chatbot

Danu Caus, John Mrziglod, Sebastian Lembcke



http://www.informatik.uni-hamburg.de/WTM/

Outline

- Motivation
- Background
- Implementation & Dataset
- Experiments & Results
- Conclusion & Improvements

Motivation

1. Project aims to create a **conversational agent** that ideally would tackle unstructured conversations, i.e. chats
2. The agent should be ideally fit for **extended** conversations with **unforeseen** topics
3. Bottom Line: Have in the end a **Chatbot** and not simply a **Dialog Agent**

"*The rules of conversation are, in general, not to dwell on any one subject, but to pass lightly from one to another without effort and without affectation; to know how to speak about trivial topics as well as serious ones*"

Encyclopedia of Diderot

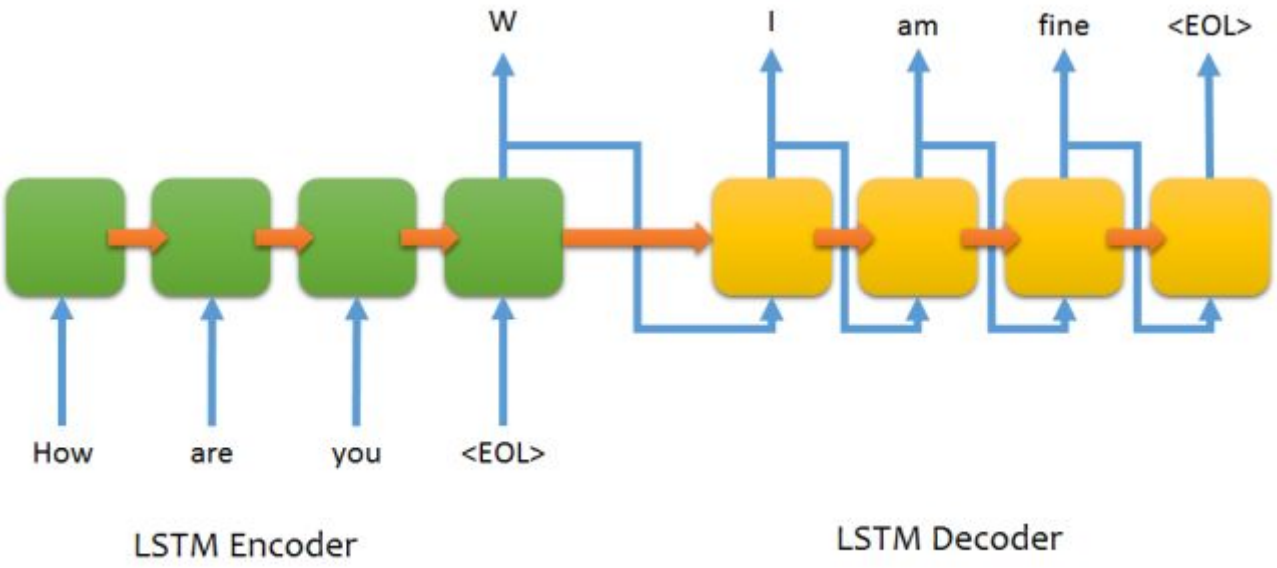
Background

- Conversational Chatbots
  - Not task oriented, i.e. not domain specific
  - Designed to simulate a human conversation partner
  - Based on turns – e.g. single word, phrase or paragraph
  - Limitation on number of turns prevents memorizing context

- Corpus-based Chatbots
  - Does not rely on handcrafted rules
  - Is trained on a dialog dataset
  - Uses machine learning or other statistical methods

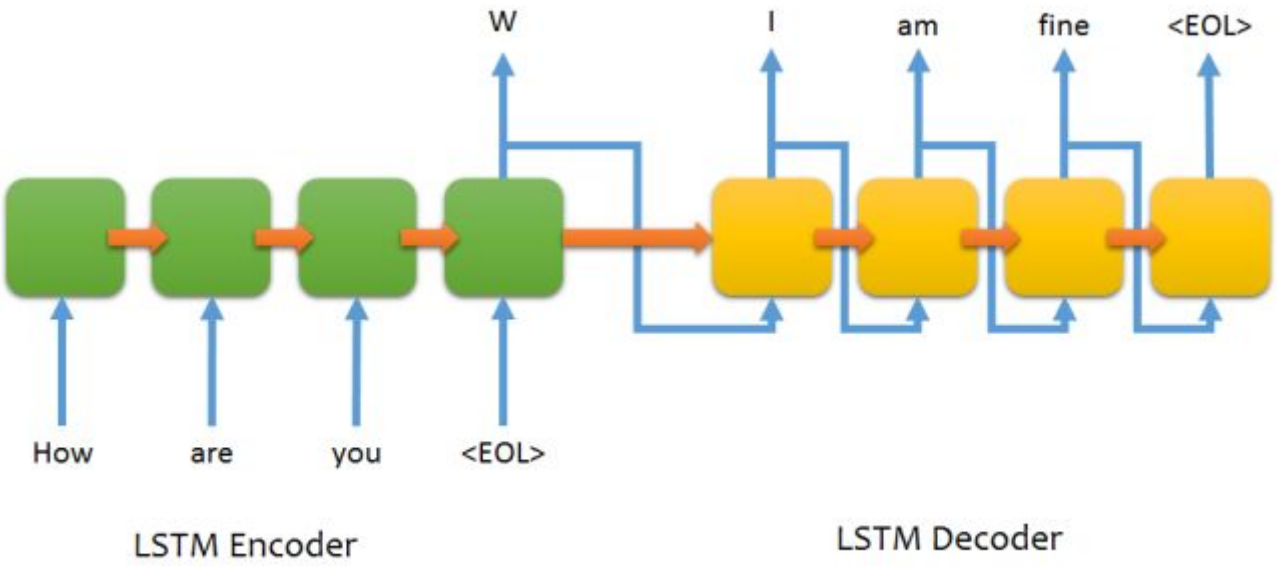
Background

- Sequence-to-Sequence Model
  - Two networks – encoder and decoder
  - Mostly recurrent neural networks (e.g. LSTM)
  - Encoder chunks sequence into words or syllables



Background

- Sequence-to-Sequence Model
  - Chunks are stored in a vocabulary
  - Encoder outputs fixed size vector
  - Vector is input for decoder, which generates new sequence



Implementation & Dataset

- Implementation
  - Tensorflow language model used for translations
  - Translation is mapping from English to English

- Dataset
  - Cornell Movie Dialogs Corpus
  - Movie lines taken from 617 Movies (304713 utterances)
  - Pre-processing to remove meaningless tags
  - Split into **train** and **test** sets

Experiments & Results

- Experiments
  - 7 experiments with vocabulary size of 40000
  - Hyperparameters:
    - Number of hidden layers 2 and 3
    - Number of hidden units 512, 800 and 1024
    - Batch size 16 and 32
  - 1 experiment with vocabulary size of 20000
    - Hyperparameters:
      - Number of hidden layers 3
      - Number of hidden units 256
      - Batch size 64

Experiments & Results

- Results

Vocabulary size 40000

Hello
fisk fisk fisk fisk fisk traded traded traded traded traded
I don't understand
particles particles particles particles ahmar ahmar ahmar ahmar ahmar ahmar traded traded traded traded traded
Hmmm what ?
grabbing fisk fisk fisk ahmar ahmar ahmar ahmar ahmar ahma

Vocabulary size 20000

and give yourself a dime
dime?
or less.
why less?
to get out of sight.
you are aggressive
not yet…

Conclusion & Improvements

- Conclusion
  - Large vocabulary may have caused underfitting
  - Result was: repeating of words and output was unrelated to input
  - Usable when trained with smaller vocabulary, yet very dramatic and unrealistic conversation
- Improvements
  - Other datasets (real conversations, not screenplays)
  - Hard coded personal information (name, age, etc.)
  - Consider previous input (remember context of conversation)
  - Ability to be trained on-the-fly by users

The End

Thank you for your attention.

Any question?

References:

[1] Fariz Rahman. Seq2Seq model.URL: https://github.com/farizrahman4u/seq2seq. Accessed Jan 29, 2018

[2] James H. Martin Daniel Jurafsky. Speech and language processing. Third Edition Draft, https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf, Accessed Jan 29 2017

[3] Cristian Danescu-Niculescu-Mizil and Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011, 2011.