The Roots of Seq2Seq Models

D. Gueorguiev 5/21/2025

The roots of the Seq2Seq Models are in a family of NLP techniques for machine translation based on sequence transduction. Let us start with ancient history and give a little theory on finite state machine transduction.

# Finite State Transducers

# Noisy Channels in Information Theory

# The Encoder-Decoder Architecture

# The Attention Mechanism

The attention scheme has been compared to the Query-Key analogy of relational databases. That

Let us consider the Attention forward pass calculating correlations of the word *“that”* with other words in *“See that girl run”*. Given the weights computed in the training process

A diagram of a machine

Description automatically generated

On the Figure the sentence is sent to three parallel streams. On the right end a single context vector emerges as a result of applying the attention mechanism to the word . The single head word embedding size is . The neuron count for each of the three subnetworks is .

denotes the row word vector (including the positional encoding) for the word .

denotes the matrix of the word embeddings of all words in the sentence. In general the dimensions of are . The attention head includes three single layer subnetworks each having neurons. The weight matrices for each of the three subnetworks are , and , all sized as .

The query component is a vector of size corresponding to single word, the key and the value matrices are sized . The function represents the softmax function. The result of applying the softmax function to is the soft weights row vector with size . Multiplying with results in the context row vector with size .

With multi-head attention we split each word vector into chunks with size .

# References

[1] [Speech and Language Processing, Daniel Jurafsky, James H. Martin, 3rd edition, Stanford U., Revised 3rd Edition, Draft, 2025](https://github.com/dimitarpg13/nlp_concepts/blob/main/literature/books/Speech_and_Language_Processing_Jurafsky_3ed_2025.pdf)

[2] [Spelling Correction and the Noisy Channel, Chapter B, Speech and Language Processing. Daniel Jurafsky & James H. Martin, Draft, 2025](https://github.com/dimitarpg13/nlp_concepts/blob/main/literature/articles/neural_machine_translation/Spelling_Correction_and_the_Noisy_Channel_Jurafsky_2024.pdf)

[3] [The Mathematics of Statistical Machine Translation: Parameter Estimation, Peter F. Brown et al, IBM, 1993](https://github.com/dimitarpg13/nlp_concepts/blob/main/literature/articles/neural_machine_translation/The_Mathematics_of_Statistical_Machine_Translation-Parameter_Estimation_Brown_1993.pdf)

[4] [An Improved Model for Noisy Channel Spelling Correction, Eric Brill and Robert C Moore, 2000](https://github.com/dimitarpg13/nlp_concepts/blob/main/literature/articles/neural_machine_translation/An_Improved_Error_Model_for_Noisy_Channel_Spelling_Correction_Brill_Moore_2000.pdf)

[5] [Introduction to Finite-State Devices in Natural Language Processing, Emmanuel Roche and Yves Schabes, 1996](https://github.com/dimitarpg13/nlp_concepts/blob/main/literature/articles/neural_machine_translation/Introduction_to_Finite-State_Devices_in_Natural_Language_Processing_roche_schabes_1996.pdf)

[6] [Finite-State Transducers in Language and Speech Processing, Mehryar Mohri, AT&T Labs Research, 1997](https://github.com/dimitarpg13/nlp_concepts/blob/main/literature/articles/neural_machine_translation/Finite-State_Transducers_in_Language_and_Speech_Processing_Mohri_1997.pdf)