IT-based Automatic Text Summarization with the Use of Textgeneration Methods

Löhr Tim

Technische Hochschule Nürnberg Georg Simon OHM

Bachelor Thesis — Business Information Systems and Management

July 27, 2020

TECHNISCHE HOCHSCHULE NÜRNBERG
GEORG SIMON OHM

Overview

1 Introduction

2 State of the Art

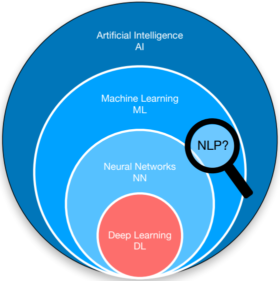
3 Prototype

4 Evaluation

Introduction

Clarify the Keywords

Artificial Intelligence is increasingly finding its way more and more into businesses.



Famous phrases for Advertisement

1. *Our product is powered now by AI!*
2. *We now use Deep Learning for a better performance!*

In conclusion: Deep Learning is a technique making use of Neural networks. Those are methods of Machine Learning, which itself is just an application of the entire AI ecosystem.

Localize my thesis within this Ecosystem

NLP

Natural Language Processing (NLP) deals with language and manipulates it to gain new information from it or perform other related tasks such as Text Summarization.

What do I use?

Using Natural Language Processing (NLP) with e.g. Hidden Markov Models is categorized as Machine Learning, whereas using Sequence Networks with e.g. the Tensorflow library is categorized Deep Learning. That is what my prototype uses.

State of the Art

Text Generation

Isn't my topic about Text Summarization? What is Text Generation?

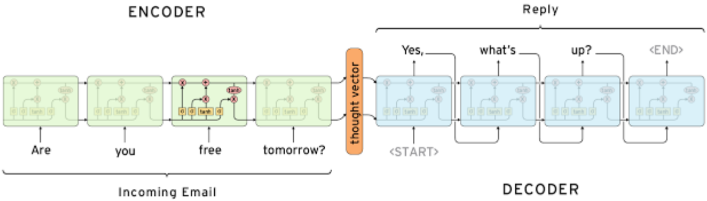


Diagram by Chris Olah

This is Text Generation performed by eight LSTM cells, which are split into four encoder and four decoder cells.

Text Generation: commonly used technologies

State of the Art in ascending order

1. Recurrent Neural Networks (RNN)
2. Long Short Term Memory (LSTM)
3. Sequence to Sequence Models
4. Encoder Decoder
5. Attention based Models

Definition Text Generation

Text Generation is a generic term for the output part of an automatic text summarizer (decoder part from the last slide). In order to understand Text Summarization, we need to know about Text Generation.

Text Summarization

To make this clear. Text Generation is the tool which produces output language, based on the preprocessed input language. Text Summarization focusses on making use of this techniques, for generating fewer words out of the original input sentence with the preferably same information content.
This is commonly known as summarization.

Today

Text Summarization in 2020 is almost not distinguishable anymore from a human summarization. A famous example is that Google uses it to automatically generate headlines for their news section, which is basically an summarization or abstraction from the news itself.

Latest Technologies in Text Summarization

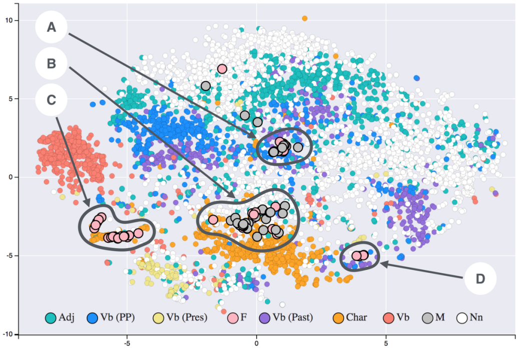
Deep Learning

Deep Learning is the best technique for Text Summarization in 2020. E.g. the famous technology *Attention*, which is based on the LSTM Neural Network, was published and open sourced by Google. Other technologies just build further up on this concept.

State of the Art in ascending order

1. Extractive approaches
2. Abstractive approaches
3. Attention
4. Pointer Generator Networks
5. Transfer Learning

How does the machine understand our language?

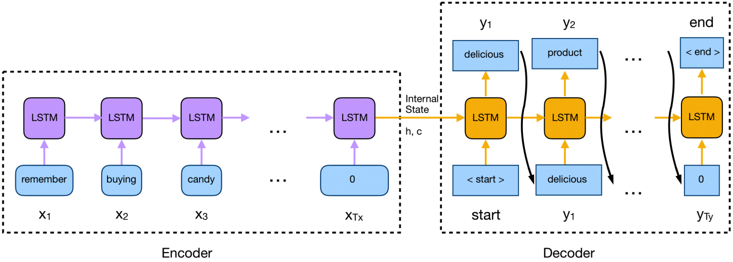


Here are shown the distribution of words by their word types in the vector space.

Well it acutally doesn't. The machine learns the structure of sentences and occurences of words in our language. In order to make them computable for the algorithm, the words will be vectorized.

Prototype

Basic structure



Architecture

My prototype is built upon multiple LSTM cells which were extended by the Attention Layer from Google. It took around 3.5 hours to train on my PC with around 222.000 training data points.

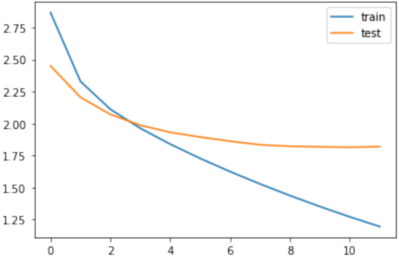
Model results

Cleaned Review	Cleaned Summary	Generated Summary
ordered chips found salty dry huge amount spices ball one bags opened	too salty and dry	too salty
found tea favorite movie theater found perfect tea guests everyone loves makes love	at the movies and home	love it
dogs special diet treats feed favorites cause problems	must be good	my dogs love these
delicious sherry flavor salad dressing great used marinade give try sweet balsamic tart red wine vinegar	yummy sweet sherry vinegar	love these
received medium roast receive correct coffee shown picture disappointed suppose ill try lot trouble return	wrong coffee received	coffee received

Evaluation

Evaluate the summary

It was necessary to clean the input text before funneling it into the model. The **cleaned summary** is the ground truth which was provided together with the dataset as labels. It can be seen, that the most important words were captures and only minor mistakes occured.



This is the Loss curve for my model from the Neural Network.