



TECHNISCHE HOCHSCHULE NÜRNBERG
GEORG SIMON OHM

Faculty of Computer Science

IT-based Text Generation with the use of NLP methods

State of the art and design of a prototype

Bachelor Thesis in
Information Systems and Management

by

Tim Löhr

Student ID 3060802

First advisor: Prof. Dr. Alfred Holl

Second advisor: Prof. Dr. Florian Gallwitz

© 2020

This work and all its parts are (protected by copyright). Any use outside the narrow limits of copyright law without the author's consent is prohibited and liable to prosecution. This applies in particular to duplications, translations, microfilming as well as storage and processing in electronic systems.

Prüfungsrechtliche Erklärung der/des Studierenden

Angaben des bzw. der Studierenden:

Name: _____ Vorname: _____ Matrikel-Nr.: _____

Fakultät: _____ Studiengang: _____

Semester: _____

Titel der Abschlussarbeit:

Ich versichere, dass ich die Arbeit selbständig verfasst, nicht anderweitig für Prüfungszwecke vorgelegt, alle benutzten Quellen und Hilfsmittel angegeben sowie wörtliche und sinngemäße Zitate als solche gekennzeichnet habe.

Ort, Datum, Unterschrift Studierende/Studierender

Erklärung zur Veröffentlichung der vorstehend bezeichneten Abschlussarbeit

Die Entscheidung über die vollständige oder auszugsweise Veröffentlichung der Abschlussarbeit liegt grundsätzlich erst einmal allein in der Zuständigkeit der/des studentischen Verfasserin/Verfassers. Nach dem Urheberrechtsgesetz (UrhG) erwirbt die Verfasserin/der Verfasser einer Abschlussarbeit mit Anfertigung ihrer/seiner Arbeit das alleinige Urheberrecht und grundsätzlich auch die hieraus resultierenden Nutzungsrechte wie z.B. Erstveröffentlichung (§ 12 UrhG), Verbreitung (§ 17 UrhG), Vervielfältigung (§ 16 UrhG), Online-Nutzung usw., also alle Rechte, die die nicht-kommerzielle oder kommerzielle Verwertung betreffen.

Die Hochschule und deren Beschäftigte werden Abschlussarbeiten oder Teile davon nicht ohne Zustimmung der/des studentischen Verfasserin/Verfassers veröffentlichen, insbesondere nicht öffentlich zugänglich in die Bibliothek der Hochschule einstellen.

Hiermit ☐ genehmige ich, wenn und soweit keine entgegenstehenden
Vereinbarungen mit Dritten getroffen worden sind,
☐ genehmige ich nicht,

dass die oben genannte Abschlussarbeit durch die Technische Hochschule Nürnberg Georg Simon Ohm, ggf. nach Ablauf einer mittels eines auf der Abschlussarbeit aufgebrachten Sperrvermerks kenntlich gemachten Sperrfrist

von _____ Jahren (0 - 5 Jahren ab Datum der Abgabe der Arbeit),

der Öffentlichkeit zugänglich gemacht wird. Im Falle der Genehmigung erfolgt diese unwiderruflich; hierzu wird der Abschlussarbeit ein Exemplar im digitalisierten PDF-Format auf einem Datenträger beigelegt. Bestimmungen der jeweils geltenden Studien- und Prüfungsordnung über Art und Umfang der im Rahmen der Arbeit abzugebenden Exemplare und Materialien werden hierdurch nicht berührt.

Ort, Datum, Unterschrift Studierende/Studierender

Preface I

The following thesis was created during the seventh and last semester at the Georg Simon Ohm University of Applied Science. Within the last three semesters, I realized that my major interest among all IT related topics is artificial intelligence.

My personal interest started basically with a group IT-project, in which my team and I programmed an autonomously driving remote control car with a deep neural network together with a Raspberry Pi 3B+. From this first project on, I selected all my further elective courses to be related to machine learning or data science in any possible way. I wanted to increase my knowledge further, so I searched for a website that provides courses related to AI. I found *www.udacity.com*, which offers courses in cooperation with top IT companies, such as Google, Airbnb, or Microsoft. Out of curiosity, I bought the course *Natural Language Processing*. After successfully finishing it, I was encouraged to write my bachelor thesis in a *Natural Language Processing* related topic. Together with my professor *Prof. Dr. Alfred Holl*, I worked out a structured methodological table for the entire structure of this paper. Even though Natural Language Processing is just a subfield of machine learning, the current state-of-the-art research is far beyond what I can research within a bachelor thesis. I decided to write my thesis about the subfield *textgeneration* within NLP. My state-of-the-art research includes all *hot topics* within NLP, and my prototype focuses only on the text generation part, to dive deeper into what NLP and especially text generation can accomplish in the year 2020.

Preface II

For my research, I encountered a lot of old and recently published papers, mostly from <https://arxiv.org/>. To read through the papers requires a lot of prior knowledge, especially in mathematics, which I learned during my semester in Hong Kong at the City University of Hong Kong. To fully understand the mathematics given in this thesis, enhanced knowledge of calculus and linear algebra is required. Even if this is not the case, I will describe the process in such a way that it can be comprehended without looking at the maths.

Machine Learning and, more specifically, NLP is not an intuitive study. I provided for the matrix notations the common terminologies originated from top researchers and tried to make the entry into this field as smooth as possible if the reader has no prior knowledge about this topic. During the five-month development process of the bachelor thesis, I gained much knowledge. I recognized that NLP is a huge topic, constantly under research. To keep up to date with the latest publications requires much effort.

To give a full state-of-the-art review about *all* NLP related disciplines is not possible within this thesis. For this reason, I focus entirely on the development of the *Neural Text Generation* (NTP), which includes more fields than the reader might imagine.

| | Titel / Kapitel | Untertitel / Unterkapitel | Woher? — Wissensinput | Wie? — Methode | Was? — Zielbeschreibung |
|---|---|---|---|---|--|
| 0 | IT-basierte Textgenerierung mit Hilfe von NLP-Methoden State of the Art & Entwurf eines Prototypen | | Allgemeinültig: • Fachbücher, Bücher • HongKong, TH-OHM • Online Kurse | 1. Was ist der State of Art von NLP - Systemen. 2. In welcher Qualität kam ich den Textgenerierungs- Prototypen selbst programmieren und welche Güte hat dieser? | 1. State of the Art fachlich herausarbeiten. 2. Einen Prototypischen Algorithmus programmieren, der zu einem gegebenen Input z.B. ein Buch immer wieder neue kreative Fortsetzungen generiert. |
| 1 | Einleitung | Fallbeispiel eines aktuellen NLP-Systems 1.1 | • [0.1] • Wissenschaftliches Schreiben und | 1. Recherche über die aktuellen und geplanten NLP-Systeme, im Bereich der Textverarbeitung. 2. Vorstellung meines Beitrags zu NLP-Systemen mithilfe meines Prototyps. | 1. Antwort auf die Frage, warum meine Bachelorarbeit sinnvoll ist und welche Motivation ich habe zur Bearbeitung 2. Erläuterung durch einen interessanten leichten Einstieg. |
| 2 | State of the Art | Relevante Aspekte der Mathematik 2.1 | [1.1] | Welches mathematische „know-how“ ist notwendig, um NLP-Systeme für Textverarbeitung und meinen Prototypen technisch verstehen zu können? | Beschreibung der anwendungsbezogenen mathematischen Modelle für diesen Themenkomplex anhand von Formeln und Erklärungen. |
| | | Geschichte des NLP 2.2 | • [0] • [0.1] | 1. Literaturrecherche über die Geschichte des NLP (40 Jahre). 2. Literaturrecherche über die ersten Einsätze der NLP-Systeme. | 1. Darstellung der Geschichte des NLP in Form einer zeitlichen Abfolge. 2. Nutzen der ersten NLP-P-Prototypen oder Technologien die im Einsatz waren. |
| | | Aktuelle Trends der Technologie 2.3 | • [0] • [0.1] • [2.2] • Fallbeispiele | 1. Literaturrecherche über aktuelle Trends (+ - 5 Jahre). 2. Recherche von aktuelle Papern und Veröffentlichungen. | 1. Darstellung der aktuellen Technologien. 2. Blick in die kurzfristige Zukunft anhand von aktuellen Fallbeispielen und Forschungsergebnissen. |
| 3 | Prototyp | Zielsetzung / Anforderungen 3.1 | • [0] • [1] • [2] | 1. Requirements Engineering. 2. Klassifizierung und Analyse möglicher Ergebnisse, z.B. ob der Output grammatikalisch korrekt ist. | 1. Erläuterung des Umfangs meines Prototyps. 2. Sammlung und Klassifizierung der Anforderungen an den Algorithmus und dessen Output. |
| | | Fachkonzept 3.2 | [3] | 1. Erstellen eines Fachkonzepts 2. Algorithmus modellieren 3. Prozessmodellierung 4. Datenflussmodellierung und, oder Datenmodellierung | 1. Fachkonzept fertig erstellt. 2. Der Prototyp wird ohne IT Bezug anhand von verschiedenen Teilmodellen modelliert. 3. Die einzelnen Prozesse werden ohne konkreten Implementierungs-Vorschlag modelliert. 4. Datenverarbeitung visualisiert |
| | | Implementierung 3.3 | [3.3] | 1. Software-Abhängigkeits-Portfolio erstellen - Vergleich geeigneter Programmiersprachen - Recherche der erforderlichen Bibliotheken - Recherche der erforderlichen Hardware, Software und Auswahl 2. Software entwickeln - Fehler reporten an Hersteller, Bib, etc. | 1. Erstellung eines IT-Konzepts in Form einer Beschreibung der notwendigen technischen Mittel anhand von Teilmodellen 2. Problemstellungen erklären und das Auftreten eines Problems „reverse Engineeren“ |
| | | Evaluation 3.4 | [3.4] | 1. Soll-Ist-Vergleich der Anforderungen mit dem Output des Prototypen. 2. Vergleich mit verwandten Arbeiten. 3. Recherche über potentielle Verbesserungen des Algorithmus. | 1. Evaluation und Analyse des Ergebnisses anhand von grammatikalischer Richtigkeit und Sinn. 2. Bessere Ergebnisse mit meinen vergleichen. 3. Optimierungsmöglichkeiten für meinen Prototypen evaluieren. |
| 4 | Generierung von übertragbarem Wissen | | [0] bis [3] | Um welche Elemente könnte mein Projekt modular erweitert werden um ein Anderes oder Besseres Ergebnis zu erzeugen und welchen Einfluss könnte es auf die Forschung haben? | Einordnung der Evaluationsergebnisse in einen gesellschaftlichen Kontext. |

Abstract

– At the end , finally finished :) –

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 1 |
| 1.1 | Structure of the thesis | 1 |
| 1.2 | Machine Learning | 2 |
| 1.3 | Case study of an Automatic Text Summarization System (ATS) | 6 |
| 2 | An Evolutionary View on the State of the Art | 9 |
| 2.1 | Structure of Text Generation | 9 |
| 2.1.1 | Text Generation Tasks | 10 |
| 2.1.2 | Architectures and Approaches | 11 |
| 2.2 | Advanced Approaches for Text Generation | 12 |
| 2.2.1 | Recurrent Neural Networks | 12 |
| 2.2.2 | Long Short Term Memory | 13 |
| 2.2.3 | Sequence to Sequence | 15 |
| 2.2.4 | Encoder and Decoder | 17 |
| 2.2.5 | Attention | 19 |
| 2.3 | Structure of Text Summarization | 20 |
| 2.3.1 | Summarization Factors | 20 |
| 2.3.2 | Input and Output Factors | 22 |
| 2.3.3 | Single Document and Multi Document | 22 |
| 2.3.4 | Extractive and Abstractive | 22 |
| 2.3.5 | Evaluation | 22 |
| 2.4 | Advanced Approaches for Text Summarization | 22 |
| 2.4.1 | Combinational Approach | 22 |
| 2.4.2 | Reinforcement Learning | 22 |
| 3 | Prototype | 23 |
| 3.1 | Objective | 23 |
| 3.2 | Technical concept | 23 |
| 3.2.1 | Structure | 23 |
| 3.2.2 | Neuronal Net | 23 |
| 3.2.3 | Process Modeling | 24 |
| 3.2.4 | Data flow modelling | 24 |
| 3.3 | Implementation | 24 |
| 3.4 | Evaluation | 24 |

| | | |
|----------|---|-----------|
| 4 | Generation of transferable knowledge | 25 |
| | List of Figures | 27 |
| | List of Tables | 29 |
| | List of Listings | 31 |
| | References | 33 |

Chapter 1

Introduction

The 21st century is flushed with an massive amount of texts and documents. Every day there are new articles, news, documentations and reports packed with information. For this reason, a new discipline, especially for students like me arose out of this. Knowledge is nowadays accessible everywhere and immediately, but the consumption takes way too much time. Websites like <https://www.blinkist.com/de> provide their costumers text summarizations of different kind of books readable in 15-30 minutes. This is an interesting way to safe time, but still this summarizations is done by hand. Artificial Intelligence researchers continuously provide knowledge to the public to summarize text with computer algorithms. The first approaches of automatic text summarization were grammatically bad and no human wants to read a grammatically broken summarizations. Bt Deep Learning changed the game completely, because algorithms are now feasible enough to summarize texts as good as humans.

1.1 Structure of the thesis

The aim of my thesis is to survey the current state of the art in text generation, especially on the focus of text summarization. The development into the state of the art neural text summarization had a huge impact on the readability for the human. For readers who are not familiar with machine learning in general, I will provide a zoom-in introduction from artificial intelligence in general into the tiny sub field text summarization. My approach is feed-forward from the definition of machine learning, deeper into the natural language processing field, further into the text generation field and within that, I focus on the text summarization part in chapter 1 - Introduction. New research and state of the art results in some natural language processing fields often lead to improvements across other related disciplines in machine learning and natural language processing, because algorithms are sometimes usable vice-versa. For this reason, I provide the most crucial text generation historical achievements in combination with the latest text summarization results, because both topics intersect in many aspects. The crucial concept of historical and modern approaches to summarize and generate text are introduced in chapter 2 - An evolutionary view on the State of the Art.



Figure 1.1: A simple Neuron with 3 inputs and 1 output [Sing 17]

To illustrate the basic workflow of a text summarizing system, I programmed a prototype. The concept, development and evaluation of this summarizer are located in chapter 4 - Prototype, but it requires prior knowledge to fully understand the mechanism from the input to the output. Finally in the last chapter I will discuss further improvements for my prototype and a brief discussing of future of text generation.

1.2 Machine Learning

In the last decade, Machine Learning (ML) is increasingly finding its way into businesses and society. Many websites and businesses use Machine Learning techniques to improve the user and costumer experience. The phrase *Machine Learning* was originally introduced in 1952 by Arthur Samuel. He developed a computer program for playing the game checkers in the 1950s. Samuel's model was based on a model of brain cell interaction by Donald Hebb from his book called *The Organization of Behavior* published in 1949. Hebb's book introduces theories on neuron excitement and the neural communication. Figure 1.1 illustrates a mathematical approximation of the humans brain cell in form of an *artificial* neuron. Nowadays, this brain-neuron based model is mostly declared to be not realistic enough [Andrew Ng, deeplearning.ai], because the structure of a brains neuron is far more complex than the illustration in figure 1.1 suggests. Nevertheless, it provides a really good entry point for this research field in my opinion.

The roots of Neural Networks (NN) lie down almost 80 years ago in 1943 when **McCulloch-Pitts** [McCu 43] compared for the first time neural networks with the structure of the human brain. The range in which Neural Networks (in the year 2020) apply to modern technologies is wide. Some disciplines have only been created due to the invention of Neural Networks, because they solve existing and new problems more effective and efficient than previously used algorithms. Many frequently held conferences around the globe proof continuous evidence

of the successes of Neural Networks. Among those various disciplines counts for example *Pattern recognition* with Convolutional Neural Networks (CNN) [Yann 98]. Convolutional Neural Networks are a one of the many special building blocks of the neural network. Every building block aims to solve a different task. For example, Pattern recognition uses different layers (building blocks) in its neural network than text summarization, because the input for Pattern recognition neural networks is often a picture consisting of e.g. 32x32 pixels, whereas the input for the text summarization is e.g. a 1000 word long text.

A widely known entry challenge into pattern recognition is the *CIFAR-10* dataset [Kriz]. It consists of 50.000 images divided into 10 classes of different objects and animals like cats and cars (5000 images of cats, 5000 images of cars, ...). Classification algorithms try now to predict a class for the input image as precise as possible. Many amateurs [Löb 19] and experts annually attempt to show their latest results in beating the former best accuracy.

Natural Language Processing is one of the various sub-fields of Machine Learning. Strictly speaking, it is actually a multidisciplinary field consisting of Artificial Intelligence (AI) and computational linguistics. Natural Language Processing is dedicated to understand and process the interactions between human (natural) language and computers. Natural Language Processing is a very broad term and can be applied on many different tasks, such as:

- **Sentiment Analysis**, e.g. Google Reviews on Restaurants
- **Machine Translation**, e.g. Google Translator
- **Speech Recognition**, e.g. Siri from Apples iPhone
- **Text Generation (Neural Text Generation [NTG])**, e.g. Text Summarization
- **Chat Bots**, e.g. Shopping Websites

Deep Learning is not an absolute definition. Many top researches define it very differently. The core can be broken down to, Deep Learning allows to build more complex neural networks, which are capable of detecting better and more correlation in data. Figure 1.2 shows the zoom-in from AI to Deep Learning. Therefore Deep Learning can be seen as a method in Machine Learning, not to mix up with Natural Language processing, which can make use of Deep Learning techniques, but it is not required to.

All of this tasks require many steps to function properly. In the broadest sense, there is always an Input and an Output, which are shown in Table 1.1.



Figure 1.2: Zoom into Artificial Intelligence from <https://rapidminer.com/blog/artificial-intelligence-machine-learning-deep-learning/>

| Example components of Input - Output systems | | | |
|--|----------------------|--------------------|----------------------|
| | Speech | Text | Images |
| Input Analysis | Speech Recognition | Text Recognition | Image Recognition |
| Output Synthesis | Generation of Speech | Generation of Text | Generation of Images |
| Processing method | NLP method | NLP method | CNN Building Blocks |

Table 1.1: A closer look into Input Output systems with the focus on Text Generation

| Examples of Natural Language Processing systems | | | |
|---|--------------|------------------|--------------------|
| | Speech | Text | Text |
| Input Analysis | Siri listens | Read in document | Read in document |
| Output Synthesis | Siri answers | Generate Summary | Generate sentiment |

Table 1.2: Examples for three different NLP tasks

It shows that Text Generation is the **output part** of a **Natural Language Processing** model. Data is collected through various different sources, e.g. images, videos or speech,

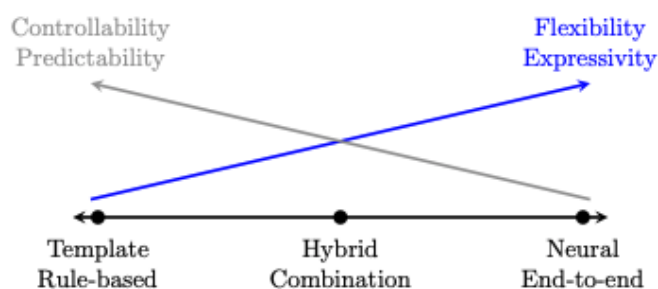


Figure 1.3: Rule-Based vs. Neural-Text-Generations System [Xie 17], Page 4

then it is further processed and generates the desired output. Useful examples are shown in Table 1.2.

For this Bachelor thesis, the focus is on the output part of a Natural Language Processing system, more specifically the text summarization, which inputs text as shown in Table 1.1 and 1.2 and outputs the summary. Text generation is therefore in general the output part of an input-output NLP system.

But what defines a summarization? Literature points out multiple different definitions. One definition proposes that the summary of a document is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks) [Mani99b]. Its objective is to give information and provide classified access to the source documents. Summarization is an automatic task, when it is generated by software in general or an algorithm.

Another term for Text Generation is *Language Modelling*, because text generators use the words of a language and grammar as input for the model. In the past five years, primarily two approaches were used for modeling a Natural Language Processing system, namely the **rule-based** system and the **template-based** system (Figure 1.3) [Xie 17]. Today neural end-to-end systems are *state-of-the-art* [Jeka 17]. These systems offer more flexibility and scale with proportionately better results, and less data is required because of the increased complexity. These systems are called neural, because they make use of **Deep Learning** Neural Networks. A major disadvantage is that the necessary computing power has increased exponentially. However, this leads to a complex problem because it becomes more and more challenging to understand the decisions of the neural network. The neural network is still, to a large extent, a *black box*. But especially in NLP it gives surprisingly good results. The neural network models for text processing are difficult to understand, so nowadays, compromises between rule-based systems still have to be made, and hybrid systems are most commonly in use.

When Neural end-to-end systems are used, Text Generation is often referred as Neural Text Generation (NTG). More examples for Neural Text Generators as output synthetical component are:

- Speech recording and conversion to text
- Conversation systems e.g. chatbots
- Neural Text summary
- Caption generation of Images

In order to train language models properly, Deep Learning (DL) algorithms teach the model the probabilities of occurring words with respect to the preceding words. There are several approaches to achieve this goal. Language models can be trained on the level of words, whole sentences, or even whole paragraphs. The granularity in which the training takes place is called *n-grams*, where *n* represents the number of preceding words. Further explanation in Section 2.2.2.1 of Chapter 2.

1.3 Case study of an Automatic Text Summarization System (ATS)

As a human, creating an good summary of a text requires that the person understood the text well. The text needs to be understood so well, that the person can summarize the texts essence in such a way, that it shortens the original document to a minimum down. However, after a period of time the person is most likely to summarize the exact same text in a different way than for example one month ago. Due to this circumstance, the summarization task tends to be challenging to automate. Depending on what kind of summary is needed, the texts must be processed in a different way, into different fragments with multiple relevance for each fragment. A crucial role is also the coherence of a text. Different applications of text summarizations are:

- Multi-Document Summarizations
- Web Page Summarization
- Reports or Meetings
- Opinion Summarizations
- Scientific Research Papers
- News Headlines

| Input: Article 1st sentence | Model-written headline |
|---|--|
| metro-goldwyn-mayer reported a third-quarter net loss of dhrs 16 million due mainly to the effect of accounting rules adopted this year | mgm reports 16 million net loss on higher revenue |
| starting from july 1, the island province of hainan in southern china will implement strict market access control on all incoming livestock and animal products to prevent the possible spread of epidemic diseases | hainan to curb spread of diseases |
| australian wine exports hit a record 52.1 million liters worth 260 million dollars (143 million us) in september, the government statistics office reported on monday | australian wine exports hit record high in september |

Table 1.3: The first column shows the first sentence of a news article which is the model input, and the second column shows what headline the model has written [Scie 15].

As an illustration and example for my entire thesis, besides of my own prototype (Chapter 3), I start with a case study for the entire Chapter 2.

Google News Headline Summarization

Google has its own news section on this link <https://news.google.com>. Google News automatically generates the news headlines for multi language news articles [Scie 15]. Google proposes that for people to digest the large amount of daily information better, they created the long-term goal at the *Google Brain* department to summarize news articles and their headlines as good as possible. The search engine Google is known for its accurate any many search results when using this engine. Google does the same for Google News. Google scrapes news articles all over the world, automatically summarizes it and out of that, it generates the headline of the summarized news article out of it [Scie 15]. The achieve their state of the art results, Google makes use of a Deep Learning technique called *sequence-to-sequence* learning, which will be explained in Section 2.2.3. Table 1.3 shows an example for generated headlines of a summarization. Due to the structure of a news article, for a good result the model who generates the headlines only needs the first few lines of an article.

All examples from this thesis will be conducted from the perspective of an news article to keep it uniform, whether a long text will be summarized to a shorter text, or only one or two sentences for a headline will be extracted and summarized from an article.

Chapter 2

An Evolutionary View on the State of the Art

The goal of this chapter is to survey the development of the text generation from the old days until 2020.

As conducted from the Introduction Chapter 1, Text Generation is the generic term for the output part of an automatic text summarizer. The research on Neural Text Generation and other fields had a major impact on the development of automatic text summarizers. In this chapter, I begin with the definition of a text generator in general and its historical development. I state out the most important steps from a basic text generator to a neural text generator. In the following I focus on the text summarizer and its personal historical evolution with the impacts of the neural text generators. There are several human-like summarizing state of the art technologies nowadays for the automatic text summarizers, but they are developed under a large scale data set and computational high demanding power. I am going to mention which parts are the fundamentals for my prototype and which technologies are state of the art, but not possible to achieve with a basic home computer.

2.1 Structure of Text Generation

Text Generation, Language modeling or Natural Language Generation are different words for basically the same meaning, but I will keep the denotation of Text Generation. A widely-cited survey from Reiter and Dale 1997 (Page 57-87) [Reit 97] characterizes Text Generation as 'the sub-field of Artificial Intelligence and computational linguistics that is concerned with the construction of computer systems than can produce understandable texts in English or other human languages from some underlying non-linguistic representation of information' [Reit 97]. This definition implies rather a data-to-text approach instead of the text-to-text approach from Table 1.1, but in 1997 the rule-based approach dominated the neural end-to-end (Neural Text Generation) methods (Figure 1.3). For that reason, in 2003 Evans declares Text Generation as quite difficult to define [Evan 02] (Page 144-151). Most researchers agree on text as the output synthesis part of the input-output system (e.g. Text Summarization or Image caption generation [Mitc 12]). Whereas the input part can be not as easily distinguished [McDo 93] (Page 191-197).

2.1.1 Text Generation Tasks

For the Text Generation input-output system, the system can be divided into six sub-problems [Reit 97]. The following bullet points contain the six most crucial steps:

- **Content determination:** Deciding which information to include in the text under construction
- **Text structuring:** Determining in which order information will be presented in the text
- **Sentence aggregation:** Deciding which information to present in individual sentences
- **Lexicalisation:** Finding the right words and phrases to express information
- **Referring expression generation:** Selecting the words and phrases to identify domain objects
- **Linguistic realization:** Combining all words and phrases into well-formed sentences

These six tasks can be thought of as early decision processes. They suggest both a chronological order in which the tasks need to be solved, as well as a distinction between strategy and tactics. This distinction goes back to Thompson H. in 1977, where he first declared these two parts [Thom 77]. Still, when it comes to modern neural state of the art Text Generation, the steps intersect in some ways. In the following comes a brief introduction to each of the steps.

2.1.1.1 Content Determination

The first step is to determine which content should be present in the generated output text. Usually there is more information stored in the input than in the output. For this reason a certain *choice* must be undertaken for the content. As mentioned in the case study (Section 1.3, the headline can be summarized very precisely given only the first few sentences of the news. In this special case the determined content could be the first three sentences. For shortening a longer document into a summary, the key points need to be abstracted into a collection of preverbal messages and semantic representations of information, often expressed in the form of a logical or database-like style [Gatt 18]. This means basically to group semantically similar words and phrases together, to remove redundancies. This step followed for the most time a rule-based approach, but in recent years researchers developed a data-driven approach (more in Section 2.1.2). For example, Barzilay and Lee (2004) developed a method to determine the content through Hidden Markov Models (HMM) [Barz 04] (Pages 113-120). Hidden Markov Models are stochastic models named after the Russian mathematician A.

A. Markow. They chain up different states of a system, in our case different topics of one or many news articles. This topics automatically will be clustered together as sentences based on the natural language semantical meaning [Gatt 18].

2.1.1.2 Text structuring

After successfully deciding which contents will be used in the generated text, the structure or order of this fragments need to be determined. Given the example article from Table 1.3:

Australian wine exports hit a record 52.1 million liters worth 260 million dollars (143 million us) in september, the government statistics office reported on monday

A good news headline should give all necessary information for the reader, namely:

- Where did it happen? -> *Australia*
- What happened? -> *Wine exports, record high*
- Who did something? -> *Australia*
- When did it happen? -> *September*

For our example the content was already predefined in the first step, now the important words and sentences will be reorder based on this four questions. Generalization approaches for the ordering task have already been proposed. Lapatas approach [Lapa 06] (Page 471-484) tries to find an optimal ordering of *information-bearing-items*. This method can even be applied to multi document input, which is more difficult to solve than single document inputs (explained in Section 2.3.3).

2.1.1.3 Sentence aggregation

2.1.1.4 Lexicalisation

2.1.1.5 Referring expression

2.1.1.6 Linguistic realisation

2.1.2 Architectures and Approaches

nlg-survey-long Kapitel 3

- Rule-based, modular approaches

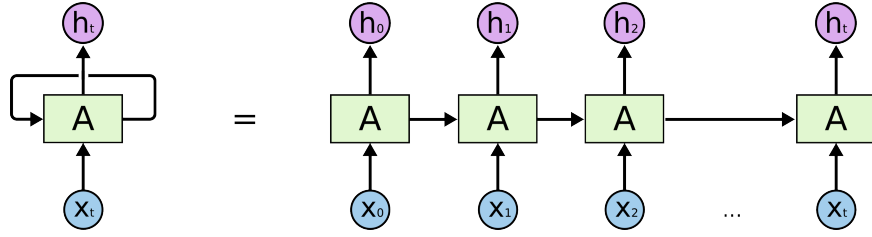


Figure 2.1: Recurrent Neural Network with integrated loops [Olah 15]

- Planning-based approaches
- Data-driven approaches

2.1.2.1 Rule-based approach

2.1.2.2 Planning-based approach

2.1.2.3 Data-driven approach

2.2 Advanced Approaches for Text Generation

2.2.1 Recurrent Neural Networks

Even though I introduced the neuron in a neural network as a kind of brain cell imitation, the neuron of a basic neural network will forget everything when it is shut down, unlike the brain. Making information persistent is a crucial step towards better performing models. Recurrent neural networks, or RNN, address this issue. They are networks with integrated loops, which allow the information to persist [Olah 15]. The network architecture of the RNN is important, because it denotes the first step into neural text generation and neural text summarization.

Figure 2.1 shows an unrolled Recurrent Neural Network. The input x_t on time step t , is passed to the neural network A . The network looks at the input on this time step and outputs the hidden state h_t at the same time step t . This loop allows the network to pass information from one time step to another. The picture 2.1 shows, that the learned parameter from input $[x]$ on time step $[t]$ will be passed as additional information to the next time step $[t + 1]$ and so on. For example if a RNN wants to predict the next word in the sentence "Since I am living in Hong Kong .. by now I speak fluent *Cantonese*". The network needs to remember that the target country is Hong Kong to predict the language Cantonese. At each time step t , the hidden state h_t of the Recurrent Neural Network is updated by:

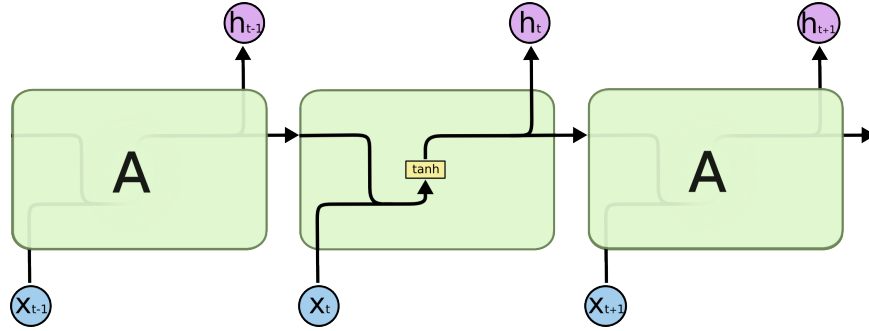


Figure 2.2: The repeating module in an Recurrent Neural Network contains one single layer [Olah 15]

$$h_{(t)} = f(h_{(t-1)}, x_t)$$

where f is a non-linear activation function and x is the input in form of a word. The function f can be in the simplest case a sigmoid function which has either 0 or 1 as output, or the more complex and effective Long Short Term Memory cell, explained in the next Section 2.2.2 [Hoch 97]. The Recurrent Neural Network is trained to predict for example the next word in a sentence or sequence. This prediction is possible due of the learned probability distribution over a sequence. The output at each time step t is a conditional distribution $p(x_t|x_{t-1}, \dots, x_1)$.

Theoretically with this approach it is possible to retain information from many time steps ago, but unfortunately, as the time span back grows, RNN's become unable to learn the information from too long ago cells. This phenomenon was explained by Sepp Hochreiter in 1991 [Hoch 91] under the name *vanishing gradient problem*. The solution to this problem is the Long Short Term Memory, short LSTM.

2.2.2 Long Short Term Memory

Long Short Term Memory cells were first proposed by Sepp Hochreiter and Jürgen Schmidhuber in 1997 [Hoch 97]. The LSTM is a special kind of Recurrent Neural Network, because it is able to remember long-term dependencies and information. The goal of the cell is to solve the vanishing gradient problem of the Recurrent Neural Network. Inputs into this cell can be stored for a long period of time, without forgetting them, as in Recurrent Neural Networks. The LSTM is designed to avoid the loss of information (vanishing gradient problem), by intentionally ledging on to certain information over plenty of time steps.

LSTM's can be enrolled the same way like RNN's, but there is a core difference between the Recurrent Neural Network in Figure 2.2 and the Long Short Term Memory in Figure 2.4. The LSTM has four gates instead of one like the RNN. The four gates are:

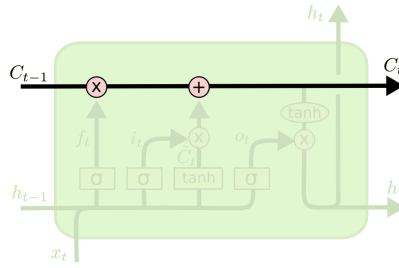


Figure 2.3: Cell State of the Long Short Term Memory which acts as data highway [Olah 15]

- Forget Gate
- Input Gate
- Cell State
- Output Gate

The **Forget Gate** decides what information should be thrown away or kept. Information from the previous hidden state and information from the current input is passed through a sigmoid function. A sigmoid function takes an input and returns high values closer to 1 and smaller values closer to 0. The closer to 0 means to forget the state, and the closer to 1 means to keep the state.

The **Input Gate** updates the cell state. That decides which values will be updated by computing the values to be between 0 and 1 like the Forget Gate. Important information is closer to 1 and 0 means less important.

The **Cell State** is the core of the LSTM. It is the horizontal line shown in Figure 2.3. The cell state acts like the information highway in the cell. With only some minor linear computation, it runs through the entire cell. This way information can pass very easily through the cell.

The **Output Gate** decides what the hidden state of the next LSTM cell should be. The hidden state contains information on previous inputs and it is also used for predictions. The hidden state denotes the state which is passed from the output gate on time step t to the input gate for the LSTM cell on time step $t+1$.

The main idea of the LSTM is, that it can decide which information to remove, to forget, which to store and when to use it. It can also decide when to move the previous state information to the next, like the RNN shown in Figure 2.1. Even though many variations of the LSTM occupy the state of the art performance, the LSTM is used in many real business cases in production, like the Google translator or weather forecasting. The Long Short Term Memory paved the way for the sequence to sequence models.

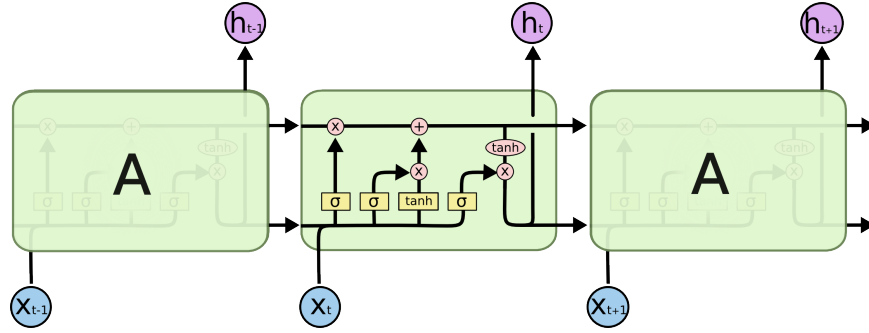


Figure 2.4: The repeating module in an LSTM contains four interacting layers [Olah 15]

2.2.2.1 N-Grams

2.2.3 Sequence to Sequence

In the year 2014, Google invented a new way to translate language by learning a statistical model with a neural machine translation approach [Suts 14]. Google called it Sequence to Sequence model [Suts 14], often shortened down to seq2seq, which consists of an encoder and a decoder.

Before that, language translation was originally processed by rule-based systems [Chen 96]. The systems computed their work by breaking down sentences into plenty of chunks and translating them phrase-by-phrase, but this approach created not easily understandable language.

After rule-based systems, statistical models have taken over the. Given a source text in e.g. German (f), what is the most suitable translation into e.g. English (e)? The statistical model $p(g|e)$ is trained on multiple texts (corpus) and finally outputs $p(e)$, which is calculated only on the target corpus in English.

$$\hat{e} = \operatorname{argmax}_e(e|g) = \operatorname{argmax}_e p(g|e)p(e)$$

The formula means, among all Bayesian probabilities $p(g|e)p(e)$, select the pair of words (translation), select the most likely to be the best translation (argmax). Even though this approach produces good results, it loses the wider semantical view, and so it is especially not effective for a good summarization technique.

For the first time, neural networks in form of feed-forward fully-connected neural networks produced such good results, that they replaced all non-network techniques. Affine matrix transformations are stacked together and are followed by non-linearities to the input and each following hidden layer [Beng 03] Page 1141-1142. However, these models require a fixed

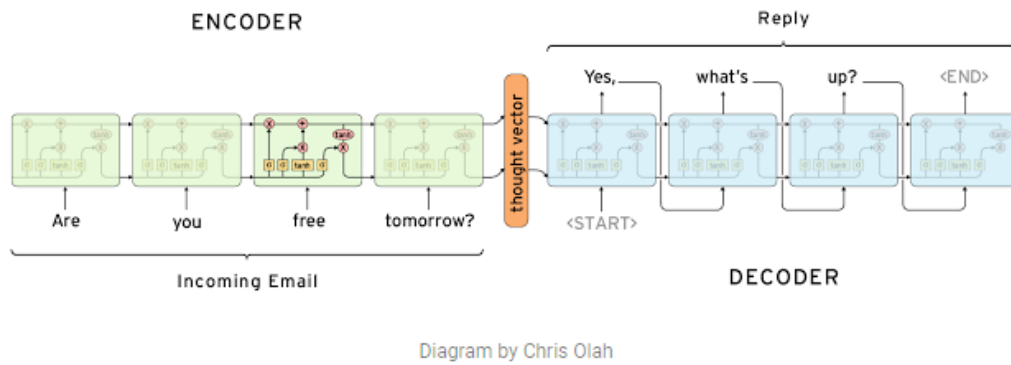


Figure 2.5: LSTM encoder-decoder model for automated E-Mail reply

content length for their calculations, which makes them again not flexible enough to produce human-like translations.

Even if a LSTM (Section 2.2.2) was used to map sequences of words from one language to another, it will most likely produce errors or bad results. A single LSTM cell needs the same input length and output length, which is unrealistic for translating multilingual. For example the English "He is running" translated into German is "Er rennt". The LSTM itself can not translate that, because of the different word length. The Long Short Term Memory cell from Section ?? was invented independently from the sequence to sequence models, but finally three employees of Google published a paper about their approach to make use of the LSTM to create a sequence to sequence model, also called encoder-decoder model. The basic idea is that the encoder converts an input text to a latent vector of length N and the decoder generates an output vector of length V by using the latent encoded vector. It is called a latent vector, because it is not accessible during the training time (manipulating it), for example in a normal Feed Forward Neural Network, the output of a hidden layer in the network can not be manipulated. The initial use of encoder-decoder models was for machine translation.

Technologies for a specific field in the machine learning environment and especially text generation can often be used cross functional. The encoder-decoder model found its way into text summarization and automated email reply by Google [Scie 15] as well. Figure 2.5 illustrates the model for Google's automated email reply.

Figure 2.5 makes use of an Long Short Term Memory cell, which captures situations, writing styles and tones. The network generalizes more flexible and accurate than a rule-based model ever could [Scie 15].

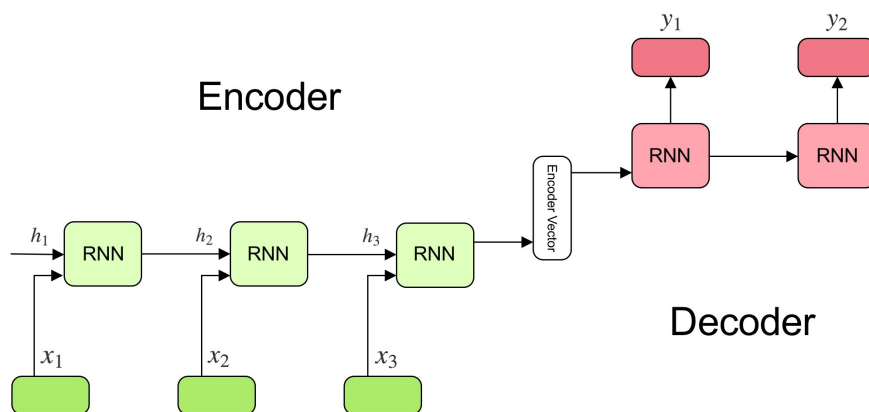


Figure 2.6: Encoder-decoder sequence to sequence model [Kost 19]

2.2.4 Encoder and Decoder

In the following, the encoder and then the decoder will be explained to have a better insight in how this technology works. The prototype from Chapter 3 is based on this kind of model. As already mentioned a sequence to sequence model is often referred to as a encoder-decoder model. The sequence to sequence model itself is built using a Recurrent Neural Network or a Long Short Term Memory as explained in the last Section 2.2.3.

Figure 2.6 shows, that the encoder decoder model is built up from actually three parts:

- Encoder
- Intermediate (encoder) Vector
- Decoder

The **Encoder** iteratively integrates the words in a sentence into the hidden state h into the Long Short Term Memory cell. Figure 2.3 shows a single LSTM cell with the input cell state C at the time step $t-1$ and the input of the hidden state h at the same time step $t-1$. This is necessary for the cell to compute both the input words, but also the knowledge from prior words. Words are represented as latent vectors in the sequence to sequence models and are stored in a vocabulary table. Each fixed length vector stands for a word in the vocabulary, for example the vector length is fixed to a dimension of 300. In a simple case, the number of words in the vocabulary is fixed to e.g. 50.000 words, hence the dimension of the vocabulary table in Figure 2.7 is [50000 x 300].

A connection of multiple recurrent units (three in Figure 2.6) where each accept a single element as an input, gains information and propagates it forward to the next cell and accordingly the next time step. In the example of Figure 2.6, the hidden state of h_3 is calculated based on the prior two cells.

| Vocabulary Table | | | | | |
|------------------|-------|-------|-------|-------|-------|
| aardvark | 1.32 | 1.56 | 0.31 | -1.21 | 0.31 |
| ate | 0.36 | -0.26 | 0.31 | -1.99 | 0.11 |
| ... | -0.69 | 0.33 | 0.77 | 0.22 | -1.29 |
| zoology | 0.41 | 0.21 | -0.32 | 0.31 | 0.22 |

(each row is actually 300 dimensions)

Figure 2.7: Snippet of an example vocabulary table [Muga 18]

The **Encoder Vector** is the last hidden state of all the encoder cells, in this example the encoder vector is located at the output of cell three. The vector tries to combine all of the information from the prior encoded words with the purpose to help the decoder make accurate predictions. Basically, the encoder vector is the initial input for the decoder part of the model.

The **Decoder** unrolls the encoder vector from meaning space into a target sentence. The meaning space (shown in Figure 2.9) is a mapping of concepts and ideas that we may want to express to points in a continuous, high-dimensional grid [Muga 18]. The minimum requirement for the meaning space is to consist at least of the last state of the encoder Recurrent Neural Network (encoder vector). The decoder computes a probability distribution for each word in the encoder vector to generate the next state. In the example case, the output is generated by multiplying the hidden state in the encoder vector h by the output matrix of size $[300 \times 50000]$. The product of this matrix multiplication is a vector of size $[50,000]$, that can be normalized with a *softmax* into a probability distribution over words in the vocabulary. The network can then choose the word with the highest probability, because the softmax squeezes all outputs into a summed up probability of 1. For example:

"Since I am living in Hong Kong, by now I speak fluent ... "

- Cat: 0.01
- running: 0.005
- Cantonese: **0.5**
- Mandarin: 0.3
- French: 0.015

The chosen word is **Cantonese**, because it is the highest probability among all probabilities which are summed up to 100%

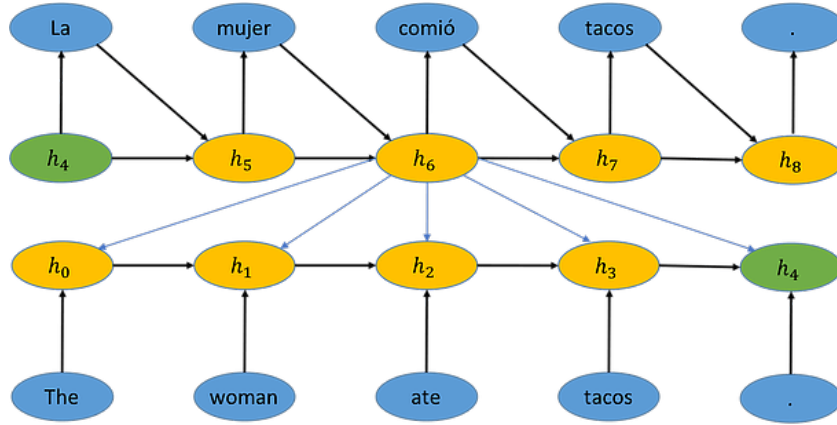


Figure 2.8: Attention mechanism for Spanish-English translation [Muga 18]

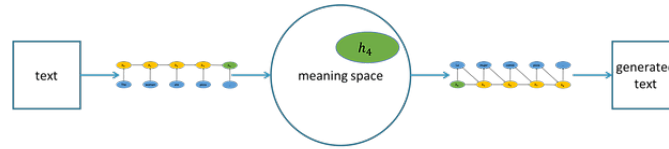


Figure 2.9: Meaning Space of the Attention model [Muga 18]

2.2.5 Attention

In general, the explanation of the sequence to sequence models just covered the very basic idea of the model. To achieve the state-of-the-art result, not only a single vector can be used for encoding the entire input sequence, but multiple vectors each capable of capturing other information.

In the encoder and decoder model, the length of the state vector h does not change for the input and output. As shown in the example of Section 2.2.3, sentences translated into another language can have a different word length. For the model to automatically adjust the length of the output, is to use the technology called *attention* [Bahd 14] [Vasw 17].

Figure 2.8 shows the basic concept of attention. The Long Short Term Memory is not starting to right before time step $t = 6$ at state h_6 . Attention enables the network to look at all prior encoded states of the words, takes the weighted average probability of the vectors and also uses this as additional information. Attention also projects its vectors into the meaning space (Figure 2.9).

Sequence to sequence models can be entirely built up from the attention model [Vasw 17].

2.2.5.1 Supervised Learning

2.2.5.2 Reinforcement Learning

2.2.5.3 GANs

2.3 Structure of Text Summarization

In the modern era of big data, retrieving useful information from a large number of textual documents is a challenging task due to the unprecedented growth in the availability of online blogs, forums, news, and scientific reports that are tremendous. Automatic text summarization provides an effective and convenient solution for reducing the amount of time it takes to read all the information. The goal of text summarization is to compress long documents into shorter summaries while maintaining the most important information and semantic of the documents [Rade 02] [Mehd 17]. Having the short summaries, the text content can be retrieved, processed and digested effectively and efficiently. Generally speaking, there are two basic approaches for performing a text summarization: Extractive and Abstractive [Mani 99a].

As mention from the Section 2.1, there is a text-to-text and data-to-text approach. At the time 2020, the transition is widely blurred, because for example the following extractive approach for text summarization is clearly a text-to-text method. It uses only the input document and nothing more. On the other side the abstractive approach is actually a data-to-text methods, because it takes into account multiple inputs, such as the text, opinion or another vocabulary.

2.3.1 Summarization Factors

Single Doc - Multi Doc Input Factors Purpose Factors output factors neural-text-summary

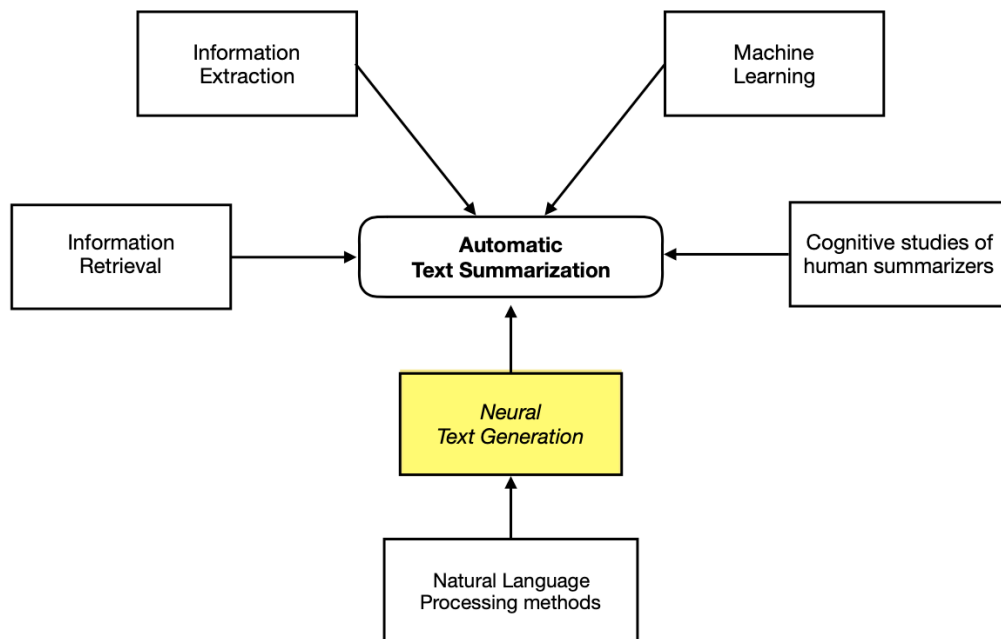


Figure 2.10: Research fields with influence on the development of text summarization

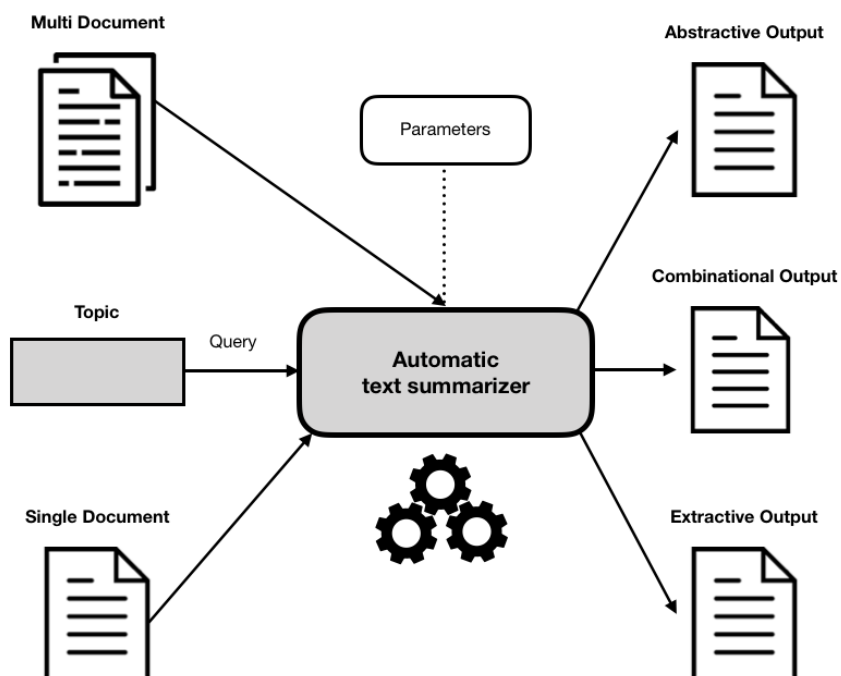


Figure 2.11: Simplified Abstraction Extraction process

2.3.2 Input and Output Factors

2.3.2.1 Input

2.3.2.2 Output

2.3.3 Single Document and Multi Document

2.3.4 Extractive and Abstractive

Figure 2.11 shows a simplified process of the abstractive or extractive approach. Both approaches can be applied decoupled from each other. The source documents, either single document or multi document go in as well with the topic of the topic (scientific, article, ...). The parameters are the compression rate τ , the type, etc.

2.3.4.1 Extractive

2.3.4.2 Abstractive

2.3.5 Evaluation

ROGUE BLEU

2.4 Advanced Approaches for Text Summarization

2.4.1 Combinational Approach

2.4.2 Reinforcement Learning

Chapter 3

Prototype

My prototype shall do this: ... , because ...

3.1 Objective

Textsummarization

3.2 Technical concept

Fachkonzept - Proto

3.2.1 Structure

The different steps of Text Generation

- Importing Dependencies
- Loading the Data
- Creating Character/Word mappings
- Data Preprocessing
- Modelling
- Generating text

3.2.2 Neuronal Net

LSTM

RNN

3.2.3 Process Modeling

NLP Bilder

3.2.4 Data flow modelling

Attention summarization

+ NLP Bilder

3.3 Implementation

Code for the Machine Translating

3.4 Evaluation

Print Ergebnisse

Bild

Image Caption

Chapter 4

Generation of transferable knowledge

Modular expandability of my project. Classification in social context

List of Figures

| | | |
|------|--|----|
| 1.1 | A simple Neuron with 3 inputs and 1 output [Sing 17] | 2 |
| 1.2 | Zoom into Artificial Intelligence from https://rapidminer.com/blog/artificial-intelligence-machine-learning-deep-learning/ | 4 |
| 1.3 | Rule-Based vs. Neural-Text-Generations System [Xie 17], Page 4 | 5 |
| 2.1 | Recurrent Neural Network with integrated loops [Olah 15] | 12 |
| 2.2 | The repeating module in an Recurrent Neural Network contains one single layer [Olah 15] | 13 |
| 2.3 | Cell State of the Long Short Term Memory which acts as data highway [Olah 15] | 14 |
| 2.4 | The repeating module in an LSTM contains four interacting layers [Olah 15] | 15 |
| 2.5 | LSTM encoder-decoder model for automated E-Mail reply | 16 |
| 2.6 | Encoder-decoder sequence to sequence model [Kost 19] | 17 |
| 2.7 | Snippet of an example vocabulary table [Muga 18] | 18 |
| 2.8 | Attention mechanism for Spanish-English translation [Muga 18] | 19 |
| 2.9 | Meaning Space of the Attention model [Muga 18] | 19 |
| 2.10 | Research fields with influence on the development of text summarization | 21 |
| 2.11 | Simplified Abstraction Extraction process | 21 |

List of Tables

| | |
|---|---|
| 1.1 A closer look into Input Output systems with the focus on Text Generation | 4 |
| 1.2 Examples for three different NLP tasks | 4 |
| 1.3 The first column shows the first sentence of a news article which is the model input, and the second column shows what headline the model has written [Scie 15]. | 7 |

List of Listings

References

- [Bahd 14] D. Bahdanau, K. Cho, and Y. Bengio. “Neural Machine Translation by Jointly Learning to Align and Translate”. 2014. cite arxiv:1409.0473Comment: Accepted at ICLR 2015 as oral presentation.
- [Barz 04] R. Barzilay and L. Lee. “Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization”. In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 113–120, Association for Computational Linguistics, Boston, Massachusetts, USA, May 2 - May 7 2004.
- [Beng 03] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. “A Neural Probabilistic Language Model”. *JOURNAL OF MACHINE LEARNING RESEARCH*, Vol. 3, pp. 1137–1155, 2003.
- [Chen 96] S. F. Chen and J. Goodman. “An Empirical Study of Smoothing Techniques for Language Modeling”. In: *34th Annual Meeting of the Association for Computational Linguistics*, pp. 310–318, Association for Computational Linguistics, Santa Cruz, California, USA, June 1996.
- [Evan 02] R. Evans, P. Piwek, and L. Cahill. “What is NLG?”. In: *Proceedings of the Second International Conference on Natural Language Generation*, pp. 144–151, Association for Computational Linguistics, 2002. Creative Commons Attribution Share-Alike License.
- [Gatt 18] A. Gatt and E. Krahmer. “Survey of the State of the Art in Natural language Generation: Core tasks, applications and evaluation”. *Journal of Artificial Intelligence Research*, Vol. 61, No. 1, pp. 65–170, 2018.
- [Hoch 91] S. Hochreiter. “Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, Institut für Informatik, Lehrstuhl Prof. Brauer, Technische Universität München”. 1991.
- [Hoch 97] S. Hochreiter and J. Schmidhuber. “Long short-term memory”. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

- [Jeka 17] O. D. Jekaterina Novikova and V. Rieser. “The E2E Dataset: New Challenges For End-to-End Generation”. 2017.
- [Kost 19] S. Kostadinov. “Understanding Encoder-Decoder Sequence to Sequence Model”. 05 2019.
- [Kriz] A. Krizhevsky, V. Nair, and G. Hinton. “CIFAR-10 (Canadian Institute for Advanced Research)”.
- [Lapa 06] M. Lapata. “Automatic Evaluation of Information Ordering: Kendall’s Tau”. *Computational Linguistics*, Vol. 32, No. 4, pp. 471–484, 2006.
- [Löh 19] T. Löhr and T. Bohnstedt. “Image Classification on the CIFAR10 Dataset”. 2019.
- [Mani 99a] I. Mani and M. Maybury. “Advances in Automatic Text Summarization”. pp. 123–136, The MIT Press, 1999.
- [Mani 99b] I. Mani. “Advances in Automatic Text Summarization (The MIT Press)”. MIT Press, 1999.
- [McCu 43] W. S. McCulloch and W. Pitts. “A logical calculus of the ideas immanent in nervous activity”. *The bulletin of mathematical biophysics*, Vol. 5, No. 4, pp. 115–133, 1943.
- [McDo 93] D. D. McDonald. “Issues in the choice of a source for Natural Language Generation”. *Computational Linguistics*, Vol. 19, No. 1, pp. 191–197, 1993.
- [Mehd 17] M. A. S. S. E. D. T. J. B. G. K. K. Mehdi Allahyari, Seyedamin Pouriyeh. “Text Summarization Techniques: A Brief Survey”. *Computation and Linguistics*, 2017.
- [Mitc 12] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, A. Mensch, A. Berg, T. Berg, and H. Daumé III. “Midge: Generating Image Descriptions From Computer Vision Detections”. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 747–756, Association for Computational Linguistics, Avignon, France, Apr. 2012.
- [Muga 18] J. Muga. “Generating Natural-Language Text with Neural Networks”. 07 2018.
- [Olah 15] C. Olah. “Understanding LSTM Networks”. 08 2015.
- [Rade 02] D. R. Radev, E. Hovy, and K. McKeown. “Introduction to the Special Issue on Summarization”. *Computational Linguistics*, Vol. 28, No. 4, pp. 399–408, 2002.
- [Reit 97] E. Reiter and R. Dale. “Building applied natural language generation systems”. *Natural Language Engineering*, Vol. 3, No. 1, pp. 57–87, 1997.

- [Scie 15] G. C. S. R. Scientist. “Computer, respond to this email.”. *Google AI Blog*, 11 2015.
- [Sing 17] P. Singh. “Neuron explained using simple algebra”. 2017.
- [Suts 14] I. Sutskever, O. Vinyals, and Q. V. Le. “Sequence to Sequence Learning with Neural Networks”. In: Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., *Advances in Neural Information Processing Systems 27*, pp. 3104–3112, Curran Associates, Inc., 2014.
- [Thom 77] H. Thompson. “Strategy and Tactics: a Model for Language Production”. *Papers from the 13th Regional Meeting of the Chicago Linguistic Society*, 1977.
- [Vasw 17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. “Attention is All you Need”. In: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., *Advances in Neural Information Processing Systems 30*, pp. 5998–6008, Curran Associates, Inc., 2017.
- [Xie 17] Z. Xie. “Neural Text Generation: A Practical Guide”. 2017.
- [Yann 98] L. B. Yann LeCun, Patrick Haffner and Y. Bengio. “Object Recognition with Gradient-Based Learning”. 1998.