Hochschule für Technik und Wirtschaft Dresden
Fakultät Informatik/ Mathematik

Abschlussarbeit zur Erlangung des akademischen Grades

Master of Science

Thema: Automatic Text Summarization
using Deep Learning and Natural Language Processing

eingereicht von: Daniel Vogel

eingereicht am: 1. Januar 2020

Erstgutachter: Prof. habil. Dr.-Ing. Hans-Joachim Böhme

Zweitgutachter: Dipl.-Kfm. Torsten Rex

Autorenreferat

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet.

Abstract

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

Formelverzeichnis

Quellcodeverzeichnis

1 Einleitung

Notizen [Backhaus et al., 2015]:

- Zweck der Automatic Text Summarization beschreiben
- Anwendungsgebiete: Report-Generierung, Nachrichten-Zusammenfassung, Überschriften-Generierung (vgl. Paper: „Automatic Text Summarization with Machine Learning“), dadurch Reduktion der Lesezeit oder auch Entscheidungsunterstützung
- Kontext und Notwendigkeit der Arbeit im Gesundheitswesen offenlegen
- Zunächst allgemeine und deutsche Texte, perspektivisch (ggf. im Ausblick erst erwähnen) fachspezifische, dialogorientierte oder auch mehrsprachige Modelle, dementsprechend notwendige Daten
- Praktischen Workflow bzw. Integration dieser Arbeit in die Praxis beschreiben
- Szenarien: Spracherkennung und Sprechererkennung auf Aufzeichnungen einer Videosprechstunde anwenden, Modelle dieser Arbeit dann für die Verdichtung der entstehenden Protokolle integrieren
- Siehe Abstract im Exposé

1.1 MediaInterface GmbH & SpeaKING®

Notizen:

- Unternehmen beschreiben
- Produkt beschreiben

1.2 Zielsetzung

Notizen:

- Ziele definieren („Automatic Single Document Summarization“, extraktiven und abstraktiven Ansatz definieren, in dieser Arbeit noch keine Zusammenfassung von Texten mit Dialogcharakter)
- Forschungsfragen formulieren

1.3 Aufbau der Arbeit

Notizen:

- Kapitel beschreiben
- Grafik als roten Faden dieser Arbeit skizzieren

1.4 Forschungsstand & Referenzen

Notizen:

- Vergleichbare Arbeiten beschreiben (vgl. Paper: „German Abstractive Text Summarization using Deep Learning“ und „Automatic Text Summarization“)
- Kürzlich entwickelte Ansätze (vgl. Paper: „Recent Automatic Text Summarization Techniques“)
- SOTA-Modelle recherchieren (vgl. Paper: „Automatic Text Summarization“ und weitere Architekturen aus dem Internet, ggf. auch ohne zugehöriges Paper)

2 Deep Learning

Notizen:

- Deep Learning definieren
- Machine Learning erwähnen
- Siehe Abstract im Exposé

2.1 Neuronale Netze

Notizen:

- Neuronale Netze definieren
- Historie beschreiben
- Funktionsweise und ausgewählte Komponenten beschreiben

2.2 Architekturen

Notizen:

- Existenz und Notwendigkeit verschiedener Architekturen ankündigen
- Später benötigte Architekturen hier beschreiben

2.2.1 MLP

Notizen:

2.2.2 RNN

Notizen:

2.2.3 LSTM

Notizen:

2.3 Hyperparameter

Notizen:

- Hyperparameter vorstellen
- Notwendigkeit und Einfluss von Hyperparametern beschreiben

3 Natural Language Processing

Notizen:

- Natural Language Processing definieren
- Umfang der Anwendungsgebiete andeuten
- Siehe Abstract im Exposé

3.1 Vorverarbeitung

Notizen:

- Pipeline der Vorverarbeitung als Voraussetzung hervorstellen

3.1.1 Textbereinigung

Notizen:

3.1.2 Tokenisierung

Notizen:

3.1.3 POS-Tagging

Notizen:

3.1.4 Lemmatisierung

Notizen:

3.1.5 Entfernen von Stoppwörtern

Notizen:

3.2 Merkmalsextraktion

Notizen:

- Relevanz für extraktiven Ansatz beschreiben (vgl. Paper: „Automatic Text Summarization“)
- Relevanz für abstraktiven Ansatz, falls vorhanden, beschreiben
- Metriken selbst weiterentwickeln und ausreifen
- Siehe: `https://scikit-learn.org/stable/modules/feature_extraction.html`

3.2.1 Übereinstimmung mit dem Titel

Notizen:

3.2.2 Satzposition

Notizen:

3.2.3 Satzähnlichkeit

Notizen:

3.2.4 Satzlänge

Notizen:

3.2.5 Domänenspezifische Wörter

Notizen:

3.2.6 Eigennamen

Notizen:

3.2.7 Numerische Daten

Notizen:

4 Datengrundlage

Notizen:

- Modelle erfordern keine gelabelten Daten, wohl aber gesichtete Daten
- Datengrundlage besteht aus frei verfügbaren allgemeinsprachlichen Daten, siehe Verzeichnis, später dann zwecks Adaption auch aus unternehmensinternen fachspezifischen Daten
- Siehe Abstract im Exposé

5 Extraktiver Ansatz

Notizen:

- Ansatz beschreiben (vgl. Paper: „Automatic Text Summarization“)
- Referenzierte Ansätze nochmal aufgreifen, kritisieren und begründen, warum meine Architektur besser ist (z.B. basierten bisherige Modelle stark auf Feature Engineering, der extraktive Ansatz ist allerdings „data-driven“ (vgl. Paper: „Extractive Text Summarization using Neural Networks“)
- Siehe Abstract im Exposé

5.1 Architektur

Notizen:

- Netzwerk des extraktiven Ansatzes als Pipeline skizzieren
- Bedingung “highly scalable” beschreiben und erfüllen (vgl. Paper: „Extractive Text Summarization using Neural Networks“)
- Sentences basierend auf Scores kopieren und ggf. neu ausrichten (vgl. Paper: „Automatic Text Summarization with Machine Learning“ und „Automatic Text Summarization Made Simple with Python“)

5.2 Konfiguration

Notizen:

5.3 Training

Notizen:

- Training verschiedener Modelle

5.4 Evaluation

Notizen:

- Kompressionsrate messen
- Qualität der Zusammenfassung messen
- Evaluation verschiedener Modelle mit geeigneter Vergleichstabelle
- Vergleich mit SOTA-Modellen
- Letztendlich soll das Modell über ein Skript berechnet werden, folglich soll eine Klasse entstehen, die einen Text und ein Modell einliest und eine Zusammenfassung ausgibt (Anforderung: inhaltlich und grammatikalisch korrekt)

6 Abstraktiver Ansatz

Notizen:

- Abgrenzung zum extraktiven Ansatz beschreiben
- Vorteile gegenüber referenzierten Modellen herausstellen
- Generierung neuer Sätze sowohl mit vorkommenden als auch mit nicht-vorkommenden Wörtern (vgl. Paper: „Automatic Text Summarization with Machine Learning“)
- Siehe Abstract im Exposé

6.1 Architektur

Notizen:

- Netzwerk des abstraktiven Ansatzes als Pipeline skizzieren

6.2 Konfiguration

Notizen:

6.3 Training

Notizen:

- Training verschiedener Modelle

6.4 Evaluation

Notizen:

- Kompressionsrate messen
- Qualität der Zusammenfassung messen
- Evaluation verschiedener Modelle mit geeigneter Vergleichstabelle

- Vergleich mit SOTA-Modellen
- Praktische Nutzung durch Implementation eines vortrainierten Modells in ein Skript oder eine Software

7 Zusammenfassung

Notizen:

- Methoden und Ergebnisse zusammenfassen
- Bewertung der Zielerreichung
- Beantwortung der Forschungsfragen
- Lösung eingangs beschriebener Szenarien
- Siehe Abstract im Exposé

8 Diskussion und Ausblick

Notizen:

- Adaptive Learning für die Modelle ansatzweise vorstellen
- Modelle für mehrere Sprachen trainieren
- Modell auf Dialogcharakter adaptieren, um es in der Verdichtung von Protokollen einer Videosprechstunde zu nutzen, bzw. generell bspw. Meetings zusammenzufassen
- Forschungsstand und SOTA-Modelle hierfür beschreiben (vgl. Paper: „Abstractive Dialogue Summarization with Sentence-Gated Modeling Optimized by Dialogue Acts“ und “Using a KG-Copy Network for Non-Goal Oriented Dialogues”), bereits Architekturen vorstellen (vgl. „Automatic Dialogue Summary Generation of Customer Service“ und „Dialogue Response Generation using Neural Networks and Background Knowledge“ und „Global Summarization of Medical Dialogue by Exploiting Local Structures”)
- Gelb markierte Literatur sichten und verwenden, Datumsangaben aktualisieren
- Siehe Abstract im Exposé

Literaturverzeichnis

[Backhaus et al., 2015] Backhaus, Klaus & Erichson, Bernd & Plinke, Wulff & Weiber, Rolf: Multivariate Analysemethoden, Verlag Springer Gabler, Berlin, Deutschland, 2015.

[Bird et al., 2009] Bird, Steven & Klein, Ewan & Loper, Edward: Natural Language Processing with Python, Verlag O'Reilly, Sebastopol, Vereinigte Staaten, 2009.

[Chaudhuri et al., 2019] Chaudhuri, Debanjan & Al Hasan Rony, Rashad & Jordan, Simon & Lehmann, Jens: Using a KG-Copy Network for Non-Goal Oriented Dialogues, Fraunhofer IAIS, Dresden, 2019.

[Gambhir et al., 2016] Gambhir, Mahak & Gupta, Vishal: Recent Automatic Text Summarization Techniques, University of Panjab in Chandigargh, 2016.

[Goncalves, 2020] Goncalves, Luis: Automatic Text Summarization with Machine Learning, in: https://medium.com/luisfredgs/automatic-text-summarization-with-machine-learning-an-overview-68ded5717a25, Aufruf am 01.01.2021.

[Goo et al., 2018] Goo, Chih-Wen & Chen, Yun-Nung: Abstractive Dialogue Summarization with Sentence-Gated Modeling Optimized by Dialogue Acts, University of Taiwan in Hsinchu, 2018.

[Joshi et al., 2020] Joshi, Anirudh & Katariya, Namit & Amatriain, Xavier & Kannan, Anitha: Global Summarization of Medical Dialogue by Exploiting Local Structures, in: `https://arxiv.org/pdf/2009.08666.pdf`, Aufruf am 01.01.2021.

[Kiani, 2017] Kiani, Farzad: Automatic Text Summarization, University of Arel in Istanbul, 2017.

[Kosovan et al., 2017] Kosovan, Sofia & Lehmann, Jens & Fischer, Asja: Dialogue Response Generation using Neural Networks with Attention and Background Knowledge, University of Bonn, 2017.

[Kriesel, 2005] Kriesel, David: Ein kleiner Überblick über neuronale Netze, in: `http://www.dkriesel.com/science/neural_networks`, Aufruf am 01.01.2021.

[Liu et al., 2019] Liu, Chunyi & Wang, Peng & Xu, Jiang & Li, Zang & Ye, Jieping: Automatic Dialogue Summary Generation for Customer Service, International Conference on Knowledge Discovery & Data Mining, Anchorage, Vereinigte Staaten, 2019.

[Nitsche, 2019] Nitsche, Matthias: Towards German Abstractive Text Summarization using Deep Learning, HAW Hamburg, 2019.

[Rashid, 2017] Rashid, Tariq: Neuronale Netze selbst programmieren, Verlag O'Reilly, Sebastopol, Vereinigte Staaten, 2017.

[Raschka et al., 2019] Raschka, Sebastian & Mirjalili, Vahid: Machine Learning and Deep Learning with Python, Verlag Packt, Birmingham, Vereinigtes Königreich, 2019.

[Sinha et al., 2018] Sinha, Akash & Abhishek, Yadav & Gahlot, Akshay: Extractive Text Summarization using Neural Networks, Indian Institute of Technology in Delhi, 2018.

[Zhang et al., 2020] Zhang, Aston & Lipton, Zachary & Li, Mu & Smola, Alexander: Dive into Deep Learning, in: `https://d2l.ai/`, Aufruf am 01.01.2021.

Thesen

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet.

Selbstständigkeitserklärung

Hiermit erkläre ich, Daniel Vogel, die vorliegende Masterarbeit selbstständig und nur unter Verwendung der von mir angegebenen Literatur verfasst zu haben. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegen.

Dresden, den 1. Januar 2021

Daniel Vogel

A Erster Anhang

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet.

B Zweiter Anhang

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet.