

**CookBERT – Adapting BERT for the Cooking Domain**

Bachelor thesis in Media Informatics at the Institute for Language, Literature and Cultural Studies (I:IMSK)

Submitted by: Pascal Strobel

Address: Gluckstraße 3, 93053 Regensburg

E-Mail (university): pascal.strobel@stud.uni-regensburg.de

E-Mail (private): paschistrobel@web.de

Matriculation number: 2106133

First reviewer: Prof. Dr. Udo Kruschwitz

Second reviewer: PD Dr. David Elsweiler

Supervisor: Alexander Frummet (M. Sc.)

Current semester: 7

Submitted on: 30.2.2016

Contents

[1 Introduction 7](#_Toc94194173)

[2 Related Work 9](#_Toc94194174)

[2.1 Transfer Learning 9](#_Toc94194175)

[2.2 Adapting BERT to specific domains 9](#_Toc94194176)

[2.3 BERT for the cooking domain 9](#_Toc94194177)

[3 BERT 9](#_Toc94194178)

[3.1 Attention 9](#_Toc94194179)

[3.2 Transformers 9](#_Toc94194180)

[3.3 Pretraining 9](#_Toc94194181)

[3.4 Finetuning 9](#_Toc94194182)

[4 Methodology 9](#_Toc94194183)

[4.1 Dataset 10](#_Toc94194184)

[4.2 DAPT/ vocabulary insertion 10](#_Toc94194185)

[4.3 Tools and environment 10](#_Toc94194186)

[4.4 Implementation details 10](#_Toc94194187)

[4.5 Evaluation tasks 10](#_Toc94194188)

[4.5.1 Multi-class classification 10](#_Toc94194189)

[4.5.2 Named-entity recognition 10](#_Toc94194190)

[4.5.3 Question Answering 10](#_Toc94194191)

[Bibliography 11](#_Toc94194192)

[Anhang A: Bausteine wissenschaftlicher Arbeiten 12](#_Toc94194193)

[A1 Theoretische Arbeit 12](#_Toc94194194)

[A2 Konstruktive Arbeit 12](#_Toc94194195)

[A3 Empirische Arbeit 12](#_Toc94194196)

[Erklärung zur Urheberschaft 14](#_Toc94194197)

Abbildungsverzeichnis (optional, in der Regel nicht notwendig)

[Abbildung 1: Blumen (Quelle, Jahr, Seitenzahl) 18](#_Toc359834290)

Tabellenverzeichnis (optional, in der Regel nicht notwendig)

[Tabelle 1: Empfohlener Textumfang 18](#_Toc359834281)

**Zusammenfassung**

Abstract

# Introduction

* [Relevanz des Themas und Motivation](https://www.scribbr.de/aufbau-und-gliederung/einleitung-beispiel/#relevanz)
* [Problembeschreibung](https://www.scribbr.de/aufbau-und-gliederung/einleitung-beispiel/#problembeschreibung)
* [Zielsetzung](https://www.scribbr.de/aufbau-und-gliederung/einleitung-beispiel/#zielsetzung)
* [Methode](https://www.scribbr.de/aufbau-und-gliederung/einleitung-beispiel/#methode)
* [Aufbau der Bachelorarbeit](https://www.scribbr.de/aufbau-und-gliederung/einleitung-beispiel/#aufbau)
* Veröffentlichung des offizielen BERT Papers (Devlin et al., 2018) vom Google AI team hat eine kleine Revolution im NLP-Bereich ausgelöst (in late 2018).
* BERT is an extremely large neural network model based on the transformer architecture (Vaswani et al., 2017), pre-trained on 3.3 Billion words (BooksCorpus 800M, Wikipedia 2.5B). / BERT is a transformer-based machine learning technique
* BERT baut auf recent work im Bereich “pre-training contextual represantations” auf – darunter Semi-supervised Sequence Learning (Dai & Le V, 2015), Generative Pre-Training (Radford et al., 2018), ELMo (Peters et al., 2018), and ULMFit (Howard & Ruder, 2018).
* Wie auch vorherige Ansätze has BERT been influenced by the new movement in NLP that is transfer learning.
* The main difference: However, unlike these previous models, BERT is the first deeply bidirectional, unsupervised language representation, pretrained using only a plain text corpus (Wikipedia and BookCorpus).
* Diese bidirektionale Eigenschaft, kombiniert mit dem Self-Attention Mechanismus ermöglicht es BERT, Wortbedeutungen und Kontext besser erfassen zu können. Das wird untermauert von den 11 downstream Tasks, in denen BERT SOTA Performance erzielt.
* Dadurch dass BERT open-sourced wurde, konnten zahlreiche Wissenschaftler weitere Benchlines Pushen und BERT für unterschiedlichste Aufgaben und Domänen, darunter Biomedical, Scientific, … oder auch Kochdomäne (A fine-tuned bidirectional encoder representations from …) anwenden mit zuversichtlichen Resultaten.
* Allerdings kann es bei domänenspezifischen Aufgaben zu Performanceeinbußen kommen, da BERT nur auf plain text data of the general text domain trainiert wurde, und somit domänenspezifisches Wissen fehlt.
* Viele Arbeiten existieren, die dieses Problem für die unterschiedlichsten Domänen in Angriff nehmen, there is however not yet a sophisticated model/ only little research done for adapting BERT to the domain of cooking.
* Ziel dieser Arbeit ist das Füllen dieser Wissenslücke von BERT via DAPT (eine gängige Methode, um BERT näher an die target domain anzunähern), indem durch das Hinzufügen von domänenspez. Wissen BERT an die Kochdomäne angenähert werden soll.

ODER:

Inspired by these latest developments (and remaining research gaps/ limitations), the goal of this bachelor thesis consists in building a sophisticated BERT model (mit mehr Daten als vorherige Ansätze) that is applicable on a variety of tasks in the cooking domain.

* Anhand einer variety of tasks (NER, MCC, QA) soll das Modell dann ausgewertet (und mit bestehenden Ansätzen verglichen) werden, um für ein möglichst breites Spektrum an NLP Aufgaben im Kochbereich anwendbar zu sein (da aktuell die verfügbaren Methoden eher eingegrenzt sind.)
* Die grundlegende Forschungsfrage, die sich daraus ergiebt, lautet:

“What are the effects of cooking domain adaption of a pre-trained BERT model on the performance of downstream tasks in this domain?”

Bzw.

“How does cooking domain adaption of a pre-trained BERT model affect the performance of downstream tasks in this domain?”

* In Kapitel 2 … 🡪 Aufbau/ Inhalte der BA erläutern

# Related Work

Hier ist der Related Work teil

## Transfer Learning

Hier transfer learning

## Adapting BERT to specific domains

Domain specific BERT models

* DAPT/ TAPT
* Pretraining from scratch

## BERT for the cooking domain

* FoodBERT und anderes cookingBERT paper + NER Paper zu FoodBase
* Datensätze enthalten häufig nur wenige Daten
* Beispiele, wann BERT im NLP angewandt wird/ werden kann:
* Foodie Fooderson
* <https://cseweb.ucsd.edu/~jmcauley/pdfs/emnlp19c.pdf>
* Conversational agents for the kitchen

# BERT

## Attention

## Transformers

Afdsfsadf

## Pretraining

## Finetuning

Dasklfjasd

# Methodology

Fasdfklsdaf

## Dataset

oadsjf

## DAPT/ vocabulary insertion

DAPT:

* If your task has a large domain-specific corpus available (e.g., "movie reviews" or "scientific papers"), it will likely be beneficial to run additional steps of pre-training on your corpus, starting from the BERT checkpoint. (https://github.com/google-research/bert#pre-training-tips-and-caveats)
* The learning rate we used in the paper was 1e-4. However, if you are doing additional steps of pre-training starting from an existing BERT checkpoint, you should use a smaller learning rate (e.g., 2e-5).

## Tools and environment

adsf

## Implementation details

asdf

## Evaluation tasks

### Multi-class classification

### Named-entity recognition

### Question Answering

# Bibliography

Anhang A: Bausteine wissenschaftlicher Arbeiten

## A1 Theoretische Arbeit

1. Fragestellung (Ziele, Motivation)
2. Überblick über Stand der Forschung und Technik (dabei Bewertung der Ansätze, Beispiele, Identifikation von Defiziten)
3. Synthese: Erstellung einer Gesamtschau (allgemeine Prinzipien, Beschreibung einer eigenen Sicht auf das Problem, Formulierung von Empfehlungen )
4. Zusammenfassung (Was wurde in der Arbeit erreicht, Erklärung des Nutzens für andere)
5. Ausblick (optional)

## A2 Konstruktive Arbeit

1. Problemstellung (Ziele, Ausgangspunkt, Vorgesehener Benutzerkreis, Bedürfnisse der Benutzer)
2. Stand der Forschung und Technik (Bisherige Lösungen, Defizite)
3. Eigenes Konzept (Lösungsansatz, allgemeines Prinzip, Werkzeuge z.B. Programmiersprachen )
4. Vorgehensweise (Beschreibung der durchgeführten Arbeitsschritte)
5. Ergebnis (Vorstellung des System z.B. Screenshots mit Erläuterungen)
6. Evaluation des System (optional, was soll evaluiert werden, welche Methode, Ablauf, Ergebnisse)
7. Zusammenfassung (Was wurde in der Arbeit erreicht; Erklärung des Nutzens für andere)
8. Ausblick (optional)

## A3 Empirische Arbeit

1. Fragestellung der Arbeit (Was soll untersucht werden, warum)
2. Stand der Forschung und Technik (Bewertung der Untersuchungs-Ansätze und Ergebnisse, Identifikation von Defiziten)
3. Präzisierung der Fragestellung (Hypothesen)
4. Untersuchungsmethodik
5. Untersuchungsablauf (Untersuchungsmaterial, Raum, Probandenrekrutierung etc.)
6. Ergebnisse (Darstellung der Ergebnisse in sinnvoller Reihenfolge, Gesamtüberblick, Einzelergebnisse z. B. geordnet nach Testcases)
7. Zusammenfassung (Was wurde erreicht, Rückbezug zu Zielen, Hypothesen, Nutzen, Erkenntnisse für weitere Untersuchungen)
8. Ausblick (optional)

Erklärung zur Urheberschaft

Ich habe die Arbeit selbständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, sowie alle Zitate und Übernahmen von fremden Aussagen kenntlich gemacht.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt.

Die vorgelegten Druckexemplare und die vorgelegte digitale Version sind identisch.

|  |  |  |
| --- | --- | --- |
| Ort, Datum |  | Unterschrift |

Erklärung zur Lizenzierung und Publikation dieser Arbeit

**Name:** Pascal Strobel

**Titel der Arbeit:** CookBERT

In der Regel räumen Sie mit Abgabe der Arbeit dem Lehrstuhl für Medieninformatik nur zwingend das Recht ein, dass die Arbeit zur Bewertung gelesen, gespeichert und vervielfältigt werden darf. Idealerweise liefern Seminararbeiten, Projektdokumentationen und Abschlussarbeiten aber einen Erkenntnisgewinn, von dem auch andere profitieren können. Wir möchten Sie deshalb bitten, uns weitere Rechte einzuräumen, bzw. idealerweise Ihre Arbeit unter eine freie Lizenz zu stellen.

Die in unseren Augen praktikabelsten Lösungen sind vorselektiert.

Hiermit gestatte ich die Verwendung der **schriftlichen Ausarbeitung** zeitlich unbegrenzt und nicht-exklusiv unter folgenden Bedingungen:

Nur zur Bewertung dieser Arbeit

Nur innerhalb des Lehrstuhls im Rahmen von Forschung und Lehre

Unter einer Creative-Commons-Lizenz mit den folgenden Einschränkungen:

BY – Namensnennung des Autors

NC – Nichtkommerziell

SA – Share-Alike, d.h. alle Änderungen müssen unter die gleiche Lizenz gestellt werden.

(An Zitaten und Abbildungen aus fremden Quellen werden keine weiteren Rechte eingeräumt.)

Außerdem gestatte ich die Verwendung des im Rahmen dieser Arbeit erstellen **Quellcodes** unter folgender Lizenz:

Nur zur Bewertung dieser Arbeit

Nur innerhalb des Lehrstuhls im Rahmen von Forschung und Lehre

Unter der CC-0-Lizenz (= beliebige Nutzung)

Unter der MIT-Lizenz (= Namensnennung)

Unter der GPLv3-Lizenz (oder neuere Versionen)

(An explizit mit einer anderen Lizenz gekennzeichneten Bibliotheken und Daten werden keine weiteren Rechte eingeräumt.)

Ich willige ein, dass der Lehrstuhl für Medieninformatik diese Arbeit – falls sie besonders gut ausfällt - auf dem Publikationsserver der Universität Regensburg veröffentlichen lässt.

Ich übertrage deshalb der Universität Regensburg das Recht, die Arbeit elektronisch zu speichern und in Datennetzen öffentlich zugänglich zu machen. Ich übertrage der Universität Regensburg ferner das Recht zur Konvertierung zum Zwecke der Langzeitarchivierung unter Beachtung der Bewahrung des Inhalts (die Originalarchivierung bleibt erhalten).

Ich erkläre außerdem, dass von mir die urheber- und lizenzrechtliche Seite (Copyright) geklärt wurde und Rechte Dritter der Publikation nicht entgegenstehen.

Ja, für die komplette Arbeit inklusive Anhang

Ja, für eine um vertrauliche Informationen gekürzte Variante (auf dem Datenträger beigefügt)

Nein

Sperrvermerk bis (Datum):

Regensburg, 10.03.2022

|  |  |  |
| --- | --- | --- |
| Ort, Datum |  | Unterschrift |

Inhalt des beigefügten Datenträgers

Beispiel (Ordner + Beschreibung):

|  |  |  |
| --- | --- | --- |
| /1\_Ausarbeitung | Die schriftliche Ausarbeitung als PDF und DOC | |
| /2\_Code | Quellcode und kompilierte Anwendung des Prototypen | |
| /3\_Studie/Design | Fragebogen und Script für die Benutzerstudie | |
| /3\_Studie/Rohdaten | Rohdaten der Studie im CSV-Format, inkl. Beschreibung der Felder | |
| /4\_Quellen | Alle in der Arbeit zitierten Quellen im PDF-Format | |
| /5\_Bilder | Alle selbst erstellten und aus anderen Quellen übernommenen Bilder | |
| /6\_Vorträge | Folien von Antritts- und Abschlussvortrag im PDF-Format | |
| /7\_Sonstiges | Notizen aus Besprechungen, Gedanken, … | |
|  | |

[Datenträger (CD, SD-Karte, o.ä.) hier oder auf Umschlaginnenseite einkleben]

Dai, A. M., & Le V, Q. (2015, November 4). *Semi-supervised Sequence Learning*. http://arxiv.org/pdf/1511.01432v1

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018, October 11). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. http://arxiv.org/pdf/1810.04805v2

Howard, J., & Ruder, S. (2018, January 18). *Universal Language Model Fine-tuning for Text Classification*. http://arxiv.org/pdf/1801.06146v5

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018, February 15). *Deep contextualized word representations*. http://arxiv.org/pdf/1802.05365v2

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). *Attention Is All You Need*. http://arxiv.org/pdf/1706.03762v5