

**CookBERT – Cooking with BERT**

**Domain-adaptive Pre-training of BERT on 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: WS 2021/22

Submitted on: 30.2.2016

Contents

[1 Introduction 8](#_Toc91705692)

[2 Ziele 10](#_Toc91705693)

[3 Related Work 10](#_Toc91705694)

[3.1 Transfer Learning 10](#_Toc91705695)

[3.2 Domain Specific BERT Models 10](#_Toc91705696)

[3.3 Adaption of Cooking Domain 10](#_Toc91705697)

[4 BERT 10](#_Toc91705698)

[4.1 Attention 10](#_Toc91705699)

[4.2 Transformers 10](#_Toc91705700)

[4.3 Pretraining 11](#_Toc91705701)

[4.4 Finetuning 11](#_Toc91705702)

[5 Methodology 11](#_Toc91705703)

[5.1 Dataset 11](#_Toc91705704)

[5.2 DAPT/ vocabulary insertion 11](#_Toc91705705)

[5.3 Tools and environment 11](#_Toc91705706)

[5.4 Implementation details 11](#_Toc91705707)

[5.5 Evaluation tasks 11](#_Toc91705708)

[5.5.1 Multi-class classification 11](#_Toc91705709)

[5.5.2 Named-entity recognition 11](#_Toc91705710)

[5.5.3 Question Answering 11](#_Toc91705711)

[Bibliography 12](#_Toc91705712)

[Anhang A: Bausteine wissenschaftlicher Arbeiten 13](#_Toc91705713)

[A1 Theoretische Arbeit 13](#_Toc91705714)

[A2 Konstruktive Arbeit 13](#_Toc91705715)

[A3 Empirische Arbeit 13](#_Toc91705716)

[Erklärung zur Urheberschaft 15](#_Toc91705717)

[Stichwortverzeichnis (optional, in der Regel nicht notwendig) 18](#_Toc91705718)

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)

Abstract

Aufgabenstellung

Die Aufgabenstellung beschreibt sowohl allgemein als auch konkret, was das Ziel der Arbeit ist, und wie dieses erreicht werden soll.

In der Regel wird die Aufgabenstellung vom Betreuer der Seminar- oder Abschlussarbeit grob vorgegeben und dann gemeinsam konkretisiert. Bei Abschlussarbeiten ist in der Regel schon eine Aufgabenstellung auf der Wiki-Seite zur Arbeit zu finden. Diese kann je nach Ausführlichkeit direkt übernommen oder als Basis für die hier dokumentierte Aufgabenstellung dienen.

Die Aufgabenstellung muss vor Anmeldung der Arbeit mit dem Betreuer abgestimmt werden. Bitte beachten Sie auch die Qualitätskriterien für Themen, die im Wiki dokumentiert sind.

Die Aufgabenstellung sollte in der Regel aus drei Abschnitten (ohne eigene Überschriften) bestehen:

* Hintergrund/Motivation
* allgemeine Zielsetzung und Herangehensweise
* konkrete einzelne Schritte zum Erreichen des Ziels

# Introduction

* Veröffentlichung von BERT hat zu einer kleinen Revolution in NLP geführt
* BERT im Vergleich zu anderen Ansätzen: “the first deeply bidirectional, unsupervised language representation, pretrained using only a plain text corpus”
* Diese Eigenschaften ermöglichen es BERT, Wortbedeutungen und Kontext besser zu erfassen und BERT erreicht in 11 vielen unterschiedlichen NLP Tasks state of the art performance 🡪 in related work dann mehrere Paper für verschiedene Aufgaben zitieren
* (Dadurch das es open source ist, konnten viele Wissenschaftler die Ergebnisse reproduzieren) 🡪 siehe Motivation von Yeung
* Das Problem: BERT wurde auf einen generellen Textkorpus vortrainiert, was zu Limitierungen von domänenspezifischen Wissen führen kann.
* Researcher haben dieses Problem schon gefaced und unterschiedliche Ansätze dagegen gefunden, darunter v.a. beliebt das Pretraining von Scratch auf domänenspez. Daten und das weitere Pretrainen eines bestehenden Modells auf
* Nachdem Ansatz 1 eine riesige Menge an Daten der Domäne und eine unglaublich hohe computational power erfordert, hat sich vor allem Ansatz zwei Etabliert (Der Ansatz führt generell, aber nicht immer, zu Verbesserung bei downstream tasks der Domäne, auf die weiter pretrained wurde)
* Mittlerweile gibt es Modelle für unterschiedlichste Domänen, darunter BioBERT, CSBERT, ConvBERT, …, allerdings gibt es aktuell noch kein sophisticated Model für die Kochdomäne.
* Vorherige Ansätze von … und … haben zwar schon derartige Modelle per DAPT trainiert, allerdings verwende ich in meinem Ansatz deutlich mehr Daten. Zudem erfolgte die Auswertung der Systeme jeweils nur anhand eines ganz spezifischen Tasks, wodurch nicht abgeleitet warden kann, ob das Modell für ein breites Spektrum an NLP tasks verwendet warden kann.
* 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.
* Die Performance wird dabei anhand einer breiten Anzahl an downstream tasks getestet und mit bereits bestehenden Modellen verglichen
* Die daraus resultierende Forschungsfrage lässt sich somit wie folgt formulieren:

“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?”

# Ziele

Die formalen und inhaltlichen Gestaltungsrichtlinien der Dokumentvorlage sollen folgenden Zielen dienen:

1. Die Vorlage soll die Erstellung formal und inhaltlich korrekter Arbeiten mit Word erleichtern.
2. Die formulierten Richtlinien können als Referenz für die Erstellung von wissenschaftlichen Arbeiten mit anderen Textverarbeitungsprogrammen dienen.

* Zu 1. Introduction hinzufügen (Ettinger, 2020)

# Related Work

Hier ist der Related Work teil

## Transfer Learning

Hier transfer learning

## Domain Specific BERT Models

Domain specific BERT models

* DAPT/ TAPT
* Pretraining from scratch

## Adaption of Cooking Domain

* FoodBERT und anderes cookingBERT paper

# BERT

## Attention

## Transformers

Afdsfsadf

## Pretraining

## Finetuning

Dasklfjasd

# Methodology

Fasdfklsdaf

## Dataset

oadsjf

## DAPT/ vocabulary insertion

dsfa

## Tools and environment

adsf

## Implementation details

asdf

## Evaluation tasks

### Multi-class classification

### Named-entity recognition

### Question Answering

# Bibliography

Ettinger, A. (2020). What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, *8*, 34–48. https://doi.org/10.1162/tacl\_a\_00298

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]