

**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 8](#_Toc94705447)

[2 Background – BERT architecture 10](#_Toc94705448)

[2.1 Attention 10](#_Toc94705449)

[2.2 Transformers 10](#_Toc94705450)

[2.3 Transfer Learning 10](#_Toc94705451)

[2.4 Pre-training 10](#_Toc94705452)

[2.5 Finetuning 10](#_Toc94705453)

[3 Related Work 11](#_Toc94705454)

[3.1 Adapting BERT to specific domains 11](#_Toc94705455)

[3.2 BERT for the cooking domain 11](#_Toc94705456)

[3.3 Summary 12](#_Toc94705457)

[4 Methodology 12](#_Toc94705458)

[4.1 Data for DAPT 12](#_Toc94705459)

[4.2 DAPT/ vocabulary insertion 12](#_Toc94705460)

[4.3 Tools and environment 12](#_Toc94705461)

[4.4 Implementation details 12](#_Toc94705462)

[4.5 Evaluation tasks 13](#_Toc94705463)

[4.5.1 Multi-class classification 13](#_Toc94705464)

[4.5.2 Named-entity recognition 13](#_Toc94705465)

[4.5.3 Question Answering 13](#_Toc94705466)

[5 Evaluation 13](#_Toc94705467)

[5.1 Multi-class classification 13](#_Toc94705468)

[5.2 Named-entity recognition 13](#_Toc94705469)

[5.3 Question Answering 13](#_Toc94705470)

[6 Discussion 13](#_Toc94705471)

[7 Limitations 13](#_Toc94705472)

[8 Conclusion and Outlook 13](#_Toc94705473)

[Bibliography 14](#_Toc94705474)

[Anhang A: Bausteine wissenschaftlicher Arbeiten 16](#_Toc94705475)

[A1 Theoretische Arbeit 16](#_Toc94705476)

[A2 Konstruktive Arbeit 16](#_Toc94705477)

[A3 Empirische Arbeit 16](#_Toc94705478)

[Erklärung zur Urheberschaft 18](#_Toc94705479)

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

* Conversational agents und converasational question answering systeme sind inzwischen ubiquitous und werden in vielen Bereichen/im everyday life eingesetzt.
* Zunehmende Verbreitung von Dialogue Systems/ Conversational agents oder auch question answering systems: basieren alle auf menschlicher Sprache/ Menschlichem Input
* One deep learning model that has established in conversational search and also in NLP in general is BERT.
* Werden in allen möglichen Domänen eingesetzt
* Zunehmend voice-based systeme
* Natürlichere Interaktion möglich
* Passend für Systeme, die einen im alltäglichen Leben begleiten (z.B. Kochen)
* Spoken queries sind anders als geschriebene: länger, mehr semantic, …
* They require a good grasp of human language. Während früher regelbasierte Systeme eingesetzt wurden, sind das heute hauptsächlich Deep Learning ansätze. Einer der neuesten Ansätze ist BERT…
* Derartige Systeme werden in allen Bereichen eingesetzt
* Conversational AI is an active research area, von rule-based systems (ELIZA, PARRY) und statistical approaches bis hin zu den heute hauptsächlichen data-driven deep neural network approaches.
* BERT ein relative neuer Ansatz, sehr gute Ergebnisse in vielen unterschiedlichen Aufgaben, da sehr gutes Sprachverständnis
* BERT inzwischen auch schon bei conversational systems eingesetzt und erzielt dabei promising results
* Ein Problem das BERT allerdings hat, is fehlendes domänenspezifisches Wissen, da eben auf generelle Textdomäne vortrainiert 🡪 Bei Küchenkonversationsassistenten begegnet BERT einer anderen Domäne 🡪 limitierung der Performance
* Allerdings gibt es noch kein existierende Model, ausgerichtet auf die Küchendomäne
* Ziel meiner Arbeit: diese Lücke füllen
* RQ: How does DAPT on cooking domain influence BERTs performance on downstream tasks that are relevant for conversational agents of this domain?
* BERT is just another step towards Google’s goal of understanding language and intent and it shouldn’t come as a surprise - especially considering new models like BERT are key for Google, due to needing to achieve higher word accuracy for voice search to avoid user frustration. (https://www.zazzlemedia.co.uk/blog/bert-update/#gref)

With the publication of Bidirectional Encoder Representations from Transformers (BERT) by the Google AI team (Devlin et al., 2018), a small revolution in the field of natural language processing (NLP) was triggered. BERT is a huge neural network model based on the transformer architecture (Vaswani et al., 2017) and was pretrained on 3.3 billion words from BooksCorpus (Zhu et al., 2015) and English Wikipedia. It builds upon previous approaches on pretraining contextual representations, particularly Semi-supervised Sequence Learning (Dai & Le, 2015), GPT (Radford et al., 2018), ELMo (Peters et al., 2018), and ULMFit (Howard & Ruder, 2018) and thus also follows the new trend in NLP: transfer learning. Transfer learning is the technique of transferring knowledge gained from performing one task to another, similar one, and is therefore beneficial compared to the resource intensive process of training neural networks (sometimes with several hundred million parameters) from scratch. However, what really sets BERT apart from other approaches is that it is “the first deeply bidirectional, unsupervised language representation, pre-trained using only a plain text corpus”, as stated in Google AI’s blog (Devlin & Chang, 2018). This bidirectionality, combined with the self-attention mechanism, provides a better grasp of word meanings and context, which is reflected in achieving state-of-the-art performance on eleven NLP tasks, as reported by Devlin et al. (2018). Through open sourcing of BERT, numerous researchers were able to utilize the model for their own tasks and domains, including the cooking domain (Stojanov et al., 2021). / Das Open Sourcing von BERT sorgt dafür, dass quasi jedermann sein eigenes Modell in kurzester Zeit für seine eigenen speziellen Aufgaben und Domänen finetunen kann, so z.B. auch für die Kochdomäne. However, BERT lacks domain-specific knowledge due to pretraining being only performed on data from the general text domain, resulting in a performance restriction on domain-specific tasks

However, BERT’s performance on domain-specific tasks is limited due to its lack of domain-specific knowledge (Gururangan et al., 2020), as pretraining was only performed on data from the general text domain. Several approaches emerged to tackle this issue, generally aiming to reduce this data shift by presenting domain-specific data of the desired domain to the model. Meanwhile, a variety of BERT models geared towards different languages and domains are available, including CamemBERT for the French language (Martin et al., 2020), BioBERT for the biomedical domain (Lee et al., 2020), SciBERT for the scientific domain (Beltagy et al., 2019), FinBERT for the financial domain (Araci, 2019), or HateBERT for abusive language (Caselli et al., 2020), just to name a few. But there is only little research done regarding the adaptation of BERT for the cooking domain, and no sophisticated model that is applicable for a broad range of downstream tasks in this domain is provided yet.

Inspired by these latest developments and remaining research gaps, the goal of this bachelor thesis consists in the adaptation of BERT for the cooking domain via domain-adaptive pretraining (DAPT), the default approach in literature to approximate a model to a specific domain. The performance of the resulting model is then evaluated against similar approaches for multiple downstream tasks, including named-entity recognition (NER), multi-class classification (MCC) and question answering (QA). This results in the underlying research question of this thesis:

*How does DAPT of BERT for the cooking domain affect the performance of downstream tasks in this domain?*

To answer this question … Aufbau erläutern

* [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)

# (Background – BERT architecture)

* BERT ist ein komplexes nn.
* Author and machine learning teacher Chris McCormick hat BERT Mountain entworfen 🡪 was muss man alles Wissen, um BERT zu verstehen 🡪 Bild von BERT mountain einfügen
* Nachdem eine Erklärung von Grund auf allerdings den Rahmen dieser Arbeit sprengen würde, werden hier nur die wichtigsten Aspekte der BERT Architektur erläutert.
  + Siehe Arbeit von Patrick Schwabl für Inhalte

# Related Work

## Adapting BERT to specific domains

* Das Anpassen von BERT an eine bestimmte Domäne ist gut untersucht und es existieren unterschiedliche Ansätze.
* Die am häufigsten vertretenen Ansätze:
  + Pretrain from scratch
  + DAPT
  + TAPT

## 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

Given the lack of exploration of neural embedding models for ingredient substitution and the challenges concerning evaluation, we propose several learningbased approaches for substitute generation and conduct both a ground truth based and a human evaluation. - FoodBERT

## Summary/ Key

Zusammenfassung von Related Work

# Methodology

Fasdfklsdaf

## Data for DAPT

* RecipeNLG

## 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

* Siehe paper von Frummet für Vorgehen (an dem orientiere ich mich eben)
* Datensatz von Frummet
* Alles so wie Frummet gemacht
* 85% train, 15% test
* No resampling
* No stopword removal
* Stratified sampling for 10 fold cross validation
* To avoid catastrophic forgetting: lower learning rate of 2e-5
* Training for 4 epochs, dropout probability of 10%, batch\_size 32
* Early stopping was included
* Wegen computing limitations wurde eine maximale Sequ. Length von 256 verwendet. D.h. wenn mehrere Turns mit angehänt wurden, wurden nur die letzten 256 tokens verwendet.
* Auswertung mit drei contexten:
  + 1. No context
    2. 1 prev turn
    3. All prev turns
* Anders als Frummet gemacht:
* Frummet hat 11 binary classifiers mit jeweils einem classificationHead der Dimension 768,2. Ich habe nur einen classifier mit classification Head mit dimension 768,11.
* Class weights were adjusted by FARMs datasilo 🡪 evlt auch machen, siehe <https://discuss.huggingface.co/t/class-weights-for-bertforsequenceclassification/1674/7>

### Named-entity recognition

### Question Answering

# Evaluation

## Multi-class classification

## Named-entity recognition

## Question Answering

# Discussion

# Limitations

# Conclusion

# Bibliography

Araci, D. (2019, August 27). *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*. http://arxiv.org/pdf/1908.10063v1

Beltagy, I., Lo Kyle, & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. http://arxiv.org/pdf/1903.10676v3

Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2020, October 23). *HateBERT: Retraining BERT for Abusive Language Detection in English*. http://arxiv.org/pdf/2010.12472v2

Dai, A. M., & Le, Q. V. (2015). Semi-supervised Sequence Learning. *Advances in Neural Information Processing Systems*, *28*, 3079–3087. http://arxiv.org/pdf/1511.01432v1

Devlin, J., & Chang, M.-W. (2018). *Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing.* Google AI. https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

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

Gururangan, S., Marasović, A., Swayamdipta, S., Lo Kyle, Beltagy, I., Downey, D., & Smith, N. A. (2020, April 23). *Don't Stop Pretraining: Adapt Language Models to Domains and Tasks*. http://arxiv.org/pdf/2004.10964v3

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

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics (Oxford, England)*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., La Clergerie, É. V. d., Seddah, D., & Sagot, B. (2020). CamemBERT: a Tasty French Language Model, 7203–7219.

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.*

Stojanov, R., Popovski, G., Cenikj, G., Koroušić Seljak, B., & Eftimov, T. (2021). A Fine-Tuned Bidirectional Encoder Representations From Transformers Model for Food Named-Entity Recognition: Algorithm Development and Validation. *Journal of Medical Internet Research*, *23*(8), e28229. https://doi.org/10.2196/28229

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

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Proceedings of the IEEE International Conference on Computer Vision*, 19–27.

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]