Exposé zur Bachelorarbeit

# Lösen des NER-Problems auf dem deutschsprachigen Onkologie Korpus BRONCO [1] mit Hilfe von Transformern

Aleksander Salek

November 2021

## 1 Motivation

Es werden immer mehr medizinische Texte publiziert und, aufgrund der Digitalisierung, auch immer mehr medizinische Daten digital gespeichert. Somit bekommt das Text-Mining auf deutsch-medizinischen Daten eine immer größer werdende Bedeutung. Text-Mining kombiniert das Information Retrieval (IR) und die Information Extraction (IE) und umfasst verschiedene Algorithmen, um unstrukturierte Daten zu analysieren. Außerdem ist der Datenschutz und die Datensicherheit ein zentrales Thema bei Text-Mining und beinhaltet die Anonymisierung und Pseudonymisierung von Patientendaten. Beispielsweise hat cloud4health[1], im deutsch-medizinischen Bereich, mit Text-Mining-Methoden ein Tool kreiert, welches u.a Texte anonymisiert oder die Plausibilität von Medikamentenverordnungen prüft [2].

IE umfasst mehrere Unteraufgaben, wovon eine die Named-Entity recognition (NER) ist. NER kann mithilfe von verschiedenen Techniken gelöst werden, unter anderem durch Transformer basierte Sprachmodelle. Mit BERT [3] wurde das erste Transformer basierte Sprachmodell vorgestellt. Dieser ermöglicht es Natural Language Processing (NLP) Tasks, z.b NER, mit höherer Genauigkeit zu lösen, als Long short-term memory (LSTM) und Conditional Random Fields (CRF), und wurde somit zu der neuen state-of-the-art Technik. Außerdem erlaubt BERT den Einsatz von transfer learning, denn man hat nun ein vortrainiertes Sprachmodell, welches auf die Bedürfnisse des Benutzers noch angepasst werden kann. In der unterstehenden Grafik (Abb. 1) [4], sieht man die NER-Performanz von BiLSTM-CRF im Vergleich zu BioBERT und MultiBERT auf vier verschiedenen Korpora, aus dem Bereich der Biomedizin. Dabei sind zwei Kopora englischsprachig und zwei auf russisch. Man kann sehen, wie beide BERT-Modelle, in Precision(P), Recall(R) und auch F1-score(F), signifikante Verbesserungen aufweisen gegenüber BiLSTM-CRF, sowohl auf den englisch- als auch auf den

[1] https://cloud4health.de/

russischsprachigen Korpora, wobei der Performanzunterschied auf den russischsprachigen Korpora noch größer ist. Somit erhoffen wir uns, dass mithilfe eines Sprachmodells, welches auf BERT basiert und weiter auf auf dem BRONCO-Korpus fine-tuned wird, die NER-Performanz in P,R und F1 gesteigert werden kann.

Der BRONCO-Korpus (BRONCO-200) besteht aus 200 manuell anonymisierten deutschen Entlassungsberichten von Krebspatienten. Zussamengefasst sind das 11.434 Sätze und 89.942 Tokens, wobei diese mit 11.124 medizinischen Entitäten und 3.118 verwandten Attributen versehen wurden. BRONCO-200 wird in zwei Teile unterteilt. Einmal in BRONCO-150, dieser besteht aus 75% des Korpusses und wird veröffentlicht. Des weiteren wurde dieser in 5 Splits unterteilt, um reproduzierbare Kreuzvalidierung durchführen zu können. Der zweite Teil des Korpusses, BRONCO-50, wird nicht veröffentlicht, um unvoreingenommene Evaluation von zukünftigen IE-tools durchführen zu können. Bisher wurde auf BRONCO-150 und BRONCO-50 NER mithilfe von CRF und LSTM (mit und ohne word embeddings) gelöst. Dabei erzielten diese auf den Annotations-Typen: „Diagnose“, „Behandlung“ und „Medikation“ eine Genauigkeit von 0.72 - 0.77, 0.81 - 0.84 und 0.88 - 0.91 im F1-score auf BRONCO-150 [1].

| **Corpus** | **Models** | **Disease** | | | **Drug** | | |
|---|---|---|---|---|---|---|---|
| | | **P** | **R** | **F** | **P** | **R** | **F** |
| EN | Multi-BERT | 55.05 | 63.91 | 59.15 | 92.21 | 92.58 | 92.39 |
| EHR | BioBERT | 56.33 | 65.56 | 60.60 | 92.39 | 92.97 | 92.68 |
| (n2c2) | LSTM-CRF | 55.00 | 56.95 | 55.96 | 89.87 | 89.70 | 89.79 |
| EN | Multi-BERT | 65.62 | 68.96 | 67.25 | 79.40 | 91.18 | 84.88 |
| UGT | BioBERT | 67.14 | 69.88 | 68.48 | 87.27 | 91.73 | 89.44 |
| (cadec) | LSTM-CRF | 64.68 | 62.77 | 63.71 | 78.50 | 70.41 | 74.23 |
| RU | Multi-BERT | 45.93 | 53.33 | 49.35 | 58.85 | 62.14 | 60.45 |
| UGT | LSTM-CRF | 27.78 | 17.44 | 21.43 | 37.74 | 40.31 | 38.98 |
| RU | Multi-BERT | 78.61 | 75.96 | 77.26 | 87.18 | 82.93 | 85.00 |
| EHR | LSTM-CRF | 62.00 | 61.69 | 61.85 | 62.00 | 79.49 | 69.66 |

Abbildung 1: NER Performance von Multi-BERT, BiLSTM-CRF und Bio-BERT im Vergleich. EHR = electronic health records, UGT = user generated Text. Tabelle aus [4].

## 2 Related work

Um NER-Probleme auf domänenspezifischen und/oder sprachspezifischen Korpora genau lösen zu können, ist es von Vorteil, speziell vortrainierte Sprachmodelle zu benutzen. In den nächsten beiden Abschnitten werden einige englisch- und deutschsprachige Sprachmodelle vorgestellt.

**Englischsprachige Sprachmodelle** Im englischsprachigen Raum gibt es einige domänenspezifische BERT-Varianten, welche auf speziellen Texten fine-getuned wurden.

Beispiele sind BioBERT [5] und SciBERT [6], welche im Folgenden kurz vorgestellt werden.

BioBERT wurde, zusätzlich zum Korpus[2] auf dem BERT [3] vortrainiert wurde, mit 4.5 Mrd. Wörtern aus PubMed Abstracts und 13.5 Mrd. Wörtern aus PubMed Central Volltext-Artikel (PMC) fine-getuned. Ergebnis sind eine 0.62 % Verbesserung im F1-score in NER Tasks auf verschiedenen medizinischen Korpora[3] im Vergleich zu, zur damaligen Zeit, state-of-the-art Methoden.

SciBERT wurde von Grund neu, auf einer zufälligen Auswahl von 1.14M Papern aus dem Semantic Scholar Korpus, vortrainiert. Dieser besteht aus 18% computerwissenschaftlicher und 82% biomedizinischer Paper. Somit wurde auf insgesamt 3.17 Mrd. Wörtern traniert. SciBERT zeigt eine Verbesserung im Lösen von NER-Task auf verschiedenen medizinischen Korpora[4]. Durchschnittlich liegt diese bei 2.06% Verbesserung im F1-score im Vergleich zu BERT [3].

**Deutschsprachige Sprachmodelle** Im deutschsprachigen Raum gibt es mehrere Sprachmodelle, welche auf BERT basieren und auch mit deutschen Texten vortrainiert und/oder fine-tuned wurden. Nachfolgend werden vier kurz vorgestellt.

Ein Vertreter ist GermanBERT. Dieses wurde von Grund auf mit 6GB deutschen Wikipedia dump, 2.4GB OpenLegalData dump und 3.6GB news Artikeln vortrainiert. Dabei übertrifft GermanBERT Google's multilingual BERT in NER-Problemen auf zwei verschiedenen Korpora[5] um ~2%.

Ein anderes Beispiel ist German Medical BERT (GerMedBERT) [7], welches GermanBERT [8] als Basis hat und mit medizinischen Artikeln aus dem Web[6] fine-tuned wurde. Zur Evaluation wurde der NTS-ICD-10-Datensatz[7] verwendet. GerMedBERT zeigt eine 2.64 % Verbesserung im F1-score bei NTS-ICD-10 document classification, womit zu vermuten wäre, dass auch die NER-Performance besser sein könnte, als mit dem GermanBERT-Modell.

Außerdem gibt es XML-RoBERTa [9], welches auf dem CoNLL 2003 (German) Korpus fine-tuned wurde. Ein Punkt, wieso dieses Modell in Frage kommt, ist, dass dieses Sprachmodell ein multilinguales Modell ist, welches auf 100 Sprachen aus dem CC-100 Korpus[8] traniert wurde und zusätzlich fine-tuned wurde wie oben beschrieben. Allein ohne das Feintuning erzielt XLM-RoBERTa eine höhere Genauigkeit beim Lösen von NER-tasks auf zwei deutschsprachigen Korpora als das German BERT und Multilingual BERT cased, wie man in Abb. 2 sehen kann.

Dann gibt es noch GottBERT [11], ein weiteres vielversprechendes Modell, welches auch auf XLM-RoBERTa [9] basiert. GottBERT wurde zudem noch auf dem deutschsprachigen Teil des OSCAR (Open Super-large Crawled Aggregated corpus)[9] trainiert, welches aus 145 GB text besteht, was ca. 21.5 Mrd. Wörter sind. In Abb. 3

[2] 2.5 Mrd. Wörter aus Wikipedia und 0.8 Mrd. Wörter aus BooksCorpus
[3] Verwendete Korpora: NCBI disease, 2010 i2b2/VA, BC5CDR u.w.
[4] Verwendete Korpora: NCBI disease, BC5CDR und JNLPBA
[5] germEval14 und CONLL03
[6] Verwendete Webseiten: netdoktor, doktoweigl, onmeda u.w.
[7] https://www.openagrar.de/receive/openagrar_mods_00046540?lang=en
[8] http://data.statmt.org/cc-100/
[9] https://oscar-corpus.com/post/oscar-v21-09/

| | XLM-RoBERTa Large | German BERT | Multilingual BERT cased | Previous Best | From |
|---|---|---|---|---|---|
| GermEval18 (Coarse) | **77.3** | 74.7 | 71.0 | 76.8 | TU Wien |
| GermEval14 | **87.0** | 84.0 | 83.4 | 84.7 | Flair |

Abbildung 2: F1-score der NER Performance von XLM-RoBERTa, German-BERT und Multilingual BERT cased im Vergleich. Tabelle aus [10]

kann man sehen, wie sich GottBERT im F1-score in der NER-Performanz auf zwei verschiedenen Korpora abhebt.

| **Model** | **CoNLL 2003** | **GermEval 2014** |
|---|---|---|
| GottBERT | **83.57** | **86.84** |
| dbmz BERT | 82.30 | 85.82 |
| $\text{mBERT}_{\text{cased}}$ | 81.20 | 85.39 |
| German BERT | 81.18 | 85.03 |
| XLM RoBERTa | 81.36 | 85.41 |

Abbildung 3: F1-score der NER Performance im Vergleich. [11]

# 3 Ziel

Bisher wurde NER auf BRONCO nur mit Hilfe von CRF und LSTM gelöst, sowohl mit als auch ohne deutsche (nicht biomedizinische) word embeddings. Ziel dieser Arbeit ist es, als Erweiterung zu [1], NER auf BRONCO mit einer höheren Genauigkeit zu lösen. Dazu werden drei verschiedene BERT Modelle auf BRONCO fine-tuned und miteinander verglichen.

# 4 Vorgehen

Mit Hilfe von Python und der dazugehörigen Bibliothek für Transformer[10] werden die oben genannten deutschsprachigen Sprachmodelle weiter auf BRONCO fine-tuned. Das Feintuning wird mithilfe der im Institut zur Verfügung stehenden GPU-Servern bewerkstelligt. Verwendet werden: GerMedBERT [7], XML-RoBERTa-GER [9] und GottBERT [11]. Alle Modelle sind auf hugginface[11] verfügbar und können weiter fine-tuned werden.

Zum einen wird NER auf BRONCO-150 gelöst. Dafür werden die Sprachmodelle auf BRONCO-150 5-fach kreuzvalidiert, da BRONCO-150 schon in 5 Splits aufgeteilt

[10] https://huggingface.co/transformers/
[11] https://huggingface.co/models

ist. Dabei fungiert jedes Split jeweils einmal als Testset zum lösen von NER und ist einmal im Trainingsset enthalten. Am Ende sind dann 5 NER-Lösungen vorhanden, über die der Durchschnitt gebildet wird.

Zum anderen wird NER auf BRONCO-50 gelöst. Hier wird jedes Sprachmodell auf BRONCO-150 fine-tuned, um damit NER auf BRONCO-50 zu lösen.

Am Ende werden die Ergebinsse der einzelnen Modelle untereinander und mit den bereits vorhandenen Ergebnissen aus [1] verglichen und bewertet.

# Literatur


[1] Madeleine Kittner u. a. „Annotation and initial evaluation of a large annotated German oncological corpus“. In: *JAMIA Open* 4.2 (Apr. 2021). ooab025. ISSN: 2574-2531. DOI: 10.1093/jamiaopen/ooab025. eprint: https://academic.oup.com/jamiaopen/article-pdf/4/2/ooab025/38830128/ooab025.pdf. URL: https://doi.org/10.1093/jamiaopen/ooab025.

[2] „Text-Mining ist eine vielversprechende Methode“. In: (2015). URL: https://www.tmf-ev.de/News/articleType/ArticleView/articleId/1689.aspx.

[3] Jacob Devlin u. a. „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“. In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[4] Zulfat Miftakhutdinov. „On Biomedical Named Entity Recognition: Experiments in Interlingual Transfer for Clinical and Social Media Texts“. In: März 2020.

[5] Jinhyuk Lee u. a. „BioBERT: a pre-trained biomedical language representation model for biomedical text mining“. In: *Bioinformatics* (Sep. 2019). Hrsg. von JonathanEditor Wren. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btz682. URL: http://dx.doi.org/10.1093/bioinformatics/btz682.

[6] Iz Beltagy, Arman Cohan und Kyle Lo. „SciBERT: Pretrained Contextualized Embeddings for Scientific Text“. In: *CoRR* abs/1903.10676 (2019). arXiv: 1903.10676. URL: http://arxiv.org/abs/1903.10676.

[7] Manjil Shrestha. „Development of a Language Model for Medical Domain“. masterthesis. Hochschule Rhein-Waal, 2021, S. 141.

[8] Milos Rusic u.a. „Open Sourcing German BERT Model“. In: (2019). URL: https://www.deepset.ai/german-bert.

[9] Julien Chaumond. „xlm-roberta-large-finetuned-conll03-german“. In: (2019). URL: https://huggingface.co/xlm-roberta-large-finetuned-conll03-german.

[10] Branden Chan. „XLM-RoBERTa: The alternative for non-english NLP“. In: (Jan. 2020). URL: https://medium.com/deepset-ai/xlm-roberta-the-multilingual-alternative-for-non-english-nlp-cf0b889ccbbf.

[11] Raphael Scheible u. a. „GottBERT: a pure German Language Model“. In: *CoRR* abs/2012.02110 (2020). arXiv: 2012.02110. URL: https://arxiv.org/abs/2012.02110.