Finetuning

Fine-tuning is straightforward since the selfattention mechanism in the Transformer allows BERT to model many downstream tasks— whether they involve single text or text pairs—by swapping out the appropriate inputs and outputs. For applications involving text pairs, a common pattern is to independently encode text pairs before applying bidirectional cross attention, such as Parikh et al. (2016); Seo et al. (2017). BERT instead uses the self-attention mechanism to unify these two stages, as encoding a concatenated text pair with self-attention effectively includes bidirectional cross attention between two sentences.

For each task, we simply plug in the taskspecific inputs and outputs into BERT and finetune all the parameters end-to-end. At the input, sentence A and sentence B from pre-training are analogous to (1) sentence pairs in paraphrasing, (2) hypothesis-premise pairs in entailment, (3) question-passage pairs in question answering, and

(4) a degenerate text-∅ pair in text classification or sequence tagging. At the output, the token representations are fed into an output layer for tokenlevel tasks, such as sequence tagging or question answering, and the [CLS] representation is fed into an output layer for classification, such as entailment or sentiment analysis.

Compared to pre-training, fine-tuning is relatively inexpensive. All of the results in the paper can be replicated in at most 1 hour on a single Cloud TPU, or a few hours on a GPU, starting from the exact same pre-trained model.7 We describe the task-specific details in the corresponding subsections of Section 4. More details can be found in Appendix A.5.

* Due to open sourcing BERT, many researchers could reconstruct the results and finetune BERT for their own downstream tasks (machine translation, ….)

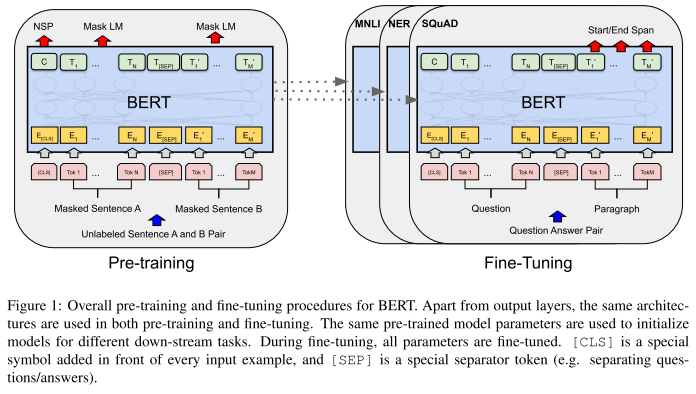


Figure 1: aus offiziellem BERT Paper

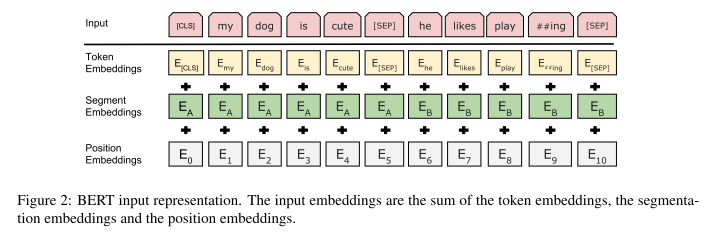


Figure 2: aus offiziellem BERT paper

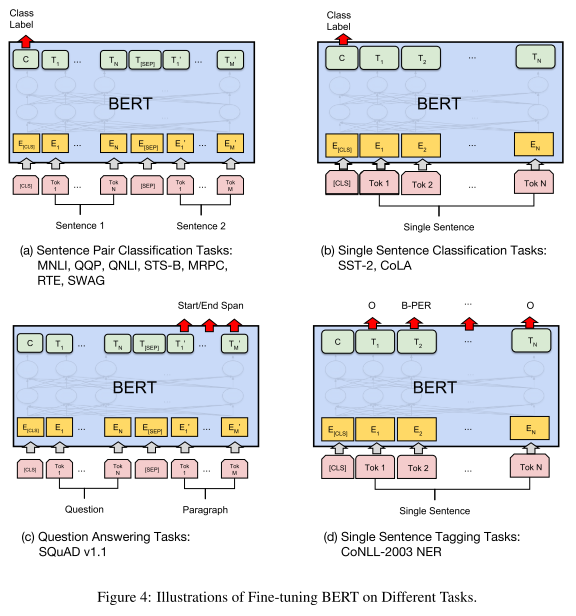


Figure 3: aus offiziellem BERT paper

# BERT and conversational agents

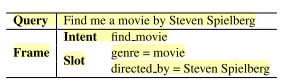
* Aufgrund der promising results of BERT for many downstream tasks, wurde es dann auch für verschiedene Aufgaben von conversational agents eingesetzt
* NLU tasks often suffer from small-scale human-labeled training data, resulting in poor generalization capability, especially for rare words. 🡪 BERT behebt das

### BERT-DST

* They present BERT-DST, a scalable ent-to-end dialogue state tracker based on BERT
* In their framework, BERT is adopted to produce contextualized representations of dialogue context, wich are used by the classification and span prediction modules to predict the slot value as none, dontcare or a text span in the dialogue context
* Dialogue state tracking: consists of determining at each turn of a dialogue the full representation of what the user wants at that point in the dialogue, which contains a goal constraint, a set of requested slots, and the user's dialogue act. The dialogue states predicted by DST are used by the downstream dialogue management component to produce API calls to a backend database and generate responses to the user. A dialogue state is often expressed as a collection of slot-value pairs (dafür ist eine ontology nötig)
* BERT-DST is an end-to-end dialogue state tracker which directly extracts slot values from the dialogue context. BERT is used as a dialogue context encoder whose contextualized language representations are suitable for scalable DST to identify slot values from their semantic contet.
* In diesem paper wird DST mit einer unknown ontology and unseen slot values behandelt
* Not requiring candidate value generation, BERTDST directly predicts slot values from the dialogue context

### BERT for joint intent classification and slot filling

* Inent class. Focuses on predicting the intent of the query, and slot filling extracts semantic concepts (slot filling = sequence labelling task, tags the input words)



* They show that BERT achieves significant improvement on intent classification accuracy, slot filling F1, and sentence-level semantic frame accuracy on several public benchmark datasets, compared to attention-based RNN models and slot-gated models.

### Fine-grained post-training for improving retrieval-based dialogue systems

* They use a fine-grained post-training method to perform better in the response selection task on three benchmark datasets (they also use a new training objective for that)
* “During the multi-turn response selection, BERT focuses on training the relationship between the context with multiple utterances and the response. However, this method of training is insufficient when considering the relations between each utterance in the context. This leads to a problem of not completely understanding the context flow that is required to select a response.”
* New Training objectives:
* Train the model by dividing the entire dialogue into multiple short context-response pairs
* Train the model with the new objective of utterance relevance classification (URC), which classifies the relation between given utterances and the target utterance into more fine-grained labels. Ähnlich zum NSP task, nurd ass hier das model unterscheidet, ob die target utterance is random or the next.

### Task-specific objectives of pre-trained language models for dialogue adaptiaton

* Pre-training auf task-indepent data enables the model zwar to learn (to some extent) universal language representations, but fails to capture crucial task-specific features
* Focus der Arbeit liegt auf Dialogue-related Natural Language Processing (DrNLP) tasks und auf dem Erstellen eines Dialogue-Adaptive Pre-training objectives (DAPO)
* Das DAPO model ist dann finetuned auf dialogue based question answering, response selection and dialogue quality evaluation 🡪 Outperformed auf allen Tasks
* Sie verwenden allerdings ELECTRA als ausgangsmodell!

### BERT with history answer embedding for conversational question answering

* One of the major challenges in multi-turn conversational search is to model the conversational history to answer the current question. Existing methods either prepend history turns to the current question or use complicated attention mechanisms to model the history. This paper proposes history answer embedding: it enables seamless itegration of conversational history into a conversational question answering model built on BERT
* Why Conversational QA? ConvQA is a simplified setting of conversational search since ConvQA systems do not focus on asking proactively, but are concrete enough for IR researchers to work on modelling the change of information needs between cycles.
* ConvQA is closely related to machine comprehension (SQUAD), nurd ass questions are organized in conversations.

### Dialogue Breakdown detection

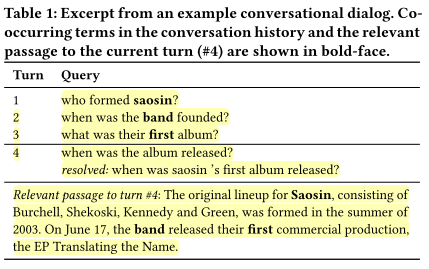
* BERT für dialogue breakdown detection
* The difficulty of developing chat-oriented dialogue systems is that such systems are required to respond to a very wide range of topics expressed by user utterances. Since it is still difficult for the current dialogue systems to continue outputting appropriate responses, utterances that cause the dialogue to collapse are often generated. It is assumed that the continuation of dialogue becomes easy when we can detect and suppress such problematic utterances

### Question rewriting for conversational question answering

* ConvQA requires the ability to correctly interpret a question in the context of previous conv turns
* Question rewriting: is specifically designed to reformulate ambiguous questions, which depend on the conversational context, into unambiguous questions that can be correctly interpreted outside of the conversational context.
* We introduce a conversational QA architecture that sets the new state of the art on the TREC CAsT 2019 passage retrieval dataset. Moreover, we show that the same QR model improves QA performance on the QuAC dataset with respect to answer span extraction, which is the next step in QA after passage retrieval. Our evaluation results indicate that the QR model we proposed achieves near human-level performance on both datasets and the gap in performance on the end-to-end conversational QA task is attributed mostly to the errors in QA
* Architektur:

Vorheriger conversational context + aktuelle implicit question = Input 🡪 Query Rewriting (GPT) 🡪 Passage Retrieval: produce ranked list of text passages from a collection, ordered by their relevance to a given natural language question. Besteht aus candidate selection (traditional retrieval algorithm BM25) und passage re-ranking (BERT). 🡪 extractive QA: given a question and a passage find the answer as a contiguous text span within the given passage (BERT)

### Query resolution for conversational search



* Query rewriting 🡪 in diesem fall: add missing context from the conversational history to the current turn query, if needed; bsp siehe Bild
* Query resolution wird in diesem Paper als binary term classification task formuliert/ vorgeschlagen (QuReTeC = Query Resolution by Term Classification): für jeden term in the previous turns of the conversation wird entschieden, ob er zu der aktuellen Query hinzugefügt werden soll oder nicht
* Das vorgeschlagene QuReTec Model basiert auf BERT: The model encodes the conversational history and the current turn query and uses a term classification layer to predict a binary label for each term in the conversational history
* When QURETEC is integrated in multi-stage ranking architecture, it significantly outperforms baseline models of different nature and is robust across conversation turns.
* Also, we found that our distant supervision method can substantially reduce the required amount of gold standard query resolutions required for training QuReTeC, using only query-passage relevance labels. This result is especially important in low resource scenarios, where gold standard query resolutions might not be readily available.

### A short survey of pretrained language models for conversational AI

* Most datasets available for NLP tasks are rather small. This data scarcity problem makes it difficult to train the deep neural networks, as they would result in an over-fitted model and not generalize well on these small datasets. The concept of pre-training a model on ImageNet corpus has been employed for quite a few years in the field of computer vision [7, 10]. The idea is to make the model learn the general features of an image and this learning can then be utilized in any vision task such as image captioning etc. to achieve the state-of-the-art results.
* The first BERT based model for QuAC was based on history answer embeddings to provide extra information to input tokens [12]. Later, [9] improved accuracy by introducing the last two contexts when answering the current question. [19] introduced the reasoning process in BERT-based architecture that improved the accuracy on the leader board drastically as compared to the previous models.
* The top positions on CoQA leaderboard 3 are occupied by pre-trained language models.
* A traditional task-oriented
* model consists of four modules namely: i) natural language understanding, ii) dialogue state tracking (DST), iii) policy learning, and iv) natural language generation.
* Lane [3] recently utilize the strengths of BERT in improving the scalability of DST module. The DST module is use to maintain the state of user’s intentions through out the dialogue. The key component ofthe model is BERT dialogue context encoding module which generates contextualized representations of the words which are very effective for mining slot values from the contextual patterns.
* In recent years, the promising notion of pre-trained language models has gained widespread attention by researchers. It is an emerging paradigm aimed to generate better contextualized representations of the words so that the dialogue systems have a better understanding of the context

### Towards topic guided conversational recommender systems (CRS)

* Conversational recommender system aims to provide high quality recommendations through conversations with users. Early CRS mainly asked questions about user preferences over predefined slots to make recommendations. Neuere Ansätze interagieren mit dem Nutzer durch natural language conversations, emphasizing fluent response generation and precise recommendation
* They use BERT for the recommendation Module: 🡪 we utilize the pretrained language model BERT to encode the historical utterancest, (+ SASRec to encode user interaction sequence). Aus diese zwei Repräsentationen kann dann die probability berechnet werden, mit der ein Item vom Item set dem user recommendet wird
* They also use BERT for the dialogue module: it predicts the next topic that guides the user tot he target topic.

Also zusammengefasst setzen viele Researchers BERT schon für conversational agents für die unterschiedilchsten Aufgaben ein. Dabei haben sie immer das standard BERT modell eingesetzt (manchmal auch angepassen BERT, der unterschieldiche pretrianing objectives hatte, bsp. History embeddings, …) was natürlich auch meistens Sinn macht, da viele der verwendeten Datensätze (wie Squad oder CoQA) auch der generellen Domäne angehört hatten. Allerdings gibt es in der Literatur sehr viele Arbeiten, die beweisen, dass BERT domänenspezifisches Wissen mangelt und daher downstream tasks gewisser Domänen limitiert sind. Das wird im nächsten Kapitel bearbeitet.

Struktur

* Conversational Agents
  + Welche Arten von Conv agents gibt es 🡪 Chatbot, Dialogue system , …
  + Conversational Agents entwicklung 🡪 von handcrafted zu neuronalen Netzen
  + Conversational Agents für welche Kontexte: Health, elderly care, …, sogar Cooking
  + Conv. Agents für die Küche:
    - Siehe <https://link.springer.com/chapter/10.1007/978-3-642-40942-4_19> fpr Mobile speech cooking assistant
    - <http://dspace.library.uvic.ca/handle/1828/9583> Caprecipes: food recommendation🡪 proof of concept für food recommender system
    - Wieso könnte das Relevant sein?
    - Es gibt mehrere Paper, die annehmen, dass das sinnvoll ist (Paper von Elsweiler usw.)
    - Frummets Paper: untersucht information needs, die beim Kochen auftreten
    - Sabrina hat Arbeit gemacht bzgl. Recommender systems
  + Conv. Agents funktionsweise/ welche NLP tasks werden benötigt/ gelöst 🡪 um dann sagen zu können, dass meine drei test-tasks relevant für conversational agents sind
    - Hier vielleicht eher allgemein sagen, dass es keine einheitliche Pipeline gibt, sondern es ganz davon abhängt, wie das System eingesetzt wird (generell gibt’s eine Komponente für Sprach-recognition, eine für Natural Language understanding, und eine für text-generation 🡪 CookBERT wäre hier für die NLU komponente sinnvoll)
* BERT Prerequisites
  + Word embeddings
  + Encoder Decoder
  + Attention
  + Transformer
  + Transfer Learning
* BERT
  + Architektur
    - Was ist anders als bei Transfomrern
    - Input darstellung
  + BERT
  + Static vs. Contextualized word embeddings
    - Was macht BERT so besonders im Vergleich zu ähnlichen Ansätzen
  + BERT Architektur
  + Pretraining
    - MLM
    - NSP
  + Finetuning
    - Für unterschiedliche Tasks, z.B. NER, Question Answering Text classification, …
    - Aufgrund der guten Performance hats auch nicht lange gedauert, bis BERT für conversational agents eingesetzt wurde:
    - Ist auch ein Konzept von Transfer Learning
  + BERT und Conversational Agents
    - Hier dann korrespondierend zum Punkt in conversational agents sagen, für welche Tasks BERT beispielsweise eingesetzt wird (und damit auch meine Tasks motivieren)
      * Intent classification
      * Answer span extraction
    - Das Problem allerdings: fehlendes Domänenspezifisches Wissen
  + Fehlendes Domänenwissen/ Domänenadaption
    - SciBERT
    - DAPT
    - Das führt dazu, dass BERT schon für unterschiedlichste Domänen verfügbar ist
  + BERT für unterschiedliche Domänen
    - BERT in der Kochdomäne
    - Hier auch schon die verschiedenen Datensätze mit reinbringen, in denen BERT angewandt wird und gute Performance erreicht (Cookversational search, DoQA, FoodBase)
* Cooking datasets relevant for conversational agents
  + Auch wenn Kochen sehr relevant ist, ist die Anzahl verfügbarer Datensätze überschaubar. Im folgenden werden 5 Datensätze vorgestellt, die in der Arbeit relevant sind
  + Die meisten Datensätze sind zudem sparse, was Transfer Learning zusätzlich motiviert
  + Die fünf Datensätze vorstellen
    - Cookversational search
    - DoQa
    - RecipeNLG
    - Recipe1M
    - FoodBase
* Summary and Key differences
  + Research Question hier reinpacken

### Cas Tasks/ Funktionsweise

* Wie funktionieren Cas, welche NLP tasks werden benötigt
* GUS:
* Framework für task-based dialogue
* First introduced in 1977 by Bobrow et al.
* Most if not all modern commercial digital asisstants are based on this
* Slot filling (= sequence labelling task und damit sehr ähnlich zu NER), Domain classification (wenn das dialogue system multi-domain ist, ansonsten nicht nötig) und intent determination (what general task or goal does the user want to accomplish?) sind für GUS Systeme relevant
* Question Answering Systems

Related work

* BERTs outstanding performance und

## BERT and Conversational Agents

* Passage re-ranking (Vakulenko et al., 2021) Answer span extraction (Vakulenko et al., 2021)
* Query rewriting (Voskarides et al., 2020)
* Sequence labelling (slot filling) (Chen et al., 2019) Intent classification (Chen et al., 2019)
* Dialog breakdown detection (Sugiyama, 2021)
* Response selection (Han et al., 2021; Wang et al., 2021)
* Dialogue state tracking (Chao & Lane, 2019)
* (Zhou et al., 2020) used BERT for a conversational recommender system for predicting items (probability von Items aus Liste bestimmen, die dem Nutzer vorgeschlagen werden) and predicting topic (next topic that guides the user to the target topic)
* (Yang, W. et al., 2019) BERT is used as a reader in their BERTserini approach (BERT macht answer extraction) 🡪 was also integrated into a chatbot
  + BERTs open sourcing allowed many researchers to apply BERT for their own tasks
  + Kurzer Hintergrund zu conversational agents und BERT
  + Conversational agents in der Küche
  + Verbnidung zw. Den beiden herstellen
  + Problem von BERT anteasern
* 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

## Adapting BERT for Different Domains

* See <https://arxiv.org/abs/1812.11806> “An introduction to domain adaptation and transfer learning”
* Welche Ansätze existieren in der Literatur und was sind jeweils vor und nachteile
* Ich nehme DAPT, da …
* BERT wurde inzwischen für die verschiedensten Domänen angepasst, bspw. …
* BERT models for cooking domain sin dim nächsten Kapitel zusammengefasst
* BERT for cooking domain
* Es gibt schon arbeiten, die BERT für die Kochdomäne anpassen
* FoodBERT: weniger Daten, nur für sehr spezielle Aufgabe getestet und nicht auf CAs ausgerichtet
* MenuNER (eher die Restaurant domäne)
* Summary and Key Differentiators
* BERT ist toll und wird auch schon für CAs eingesetzt
* BERT fehlt domänenspez. Wissen
* BERT gibts für viele Domänen, allerdings noch nicht wirklich für die Kochdomäne
* 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

* Auswirkungen der Datenmenge auf DAPT kurz erläutern.

## BERT for the Cooking Domain

## Summary and Key Differentiators

Andere Arbeiten haben BERT zwar schon für die Kochdomäne angepasst, allerdings mach ich 2 Sachen anders:

* Größeren Datensatz für DAPT, was laut Literatur generell zu besseren Ergebnissen führen kann
* Bisherige Modelle (FoodBERT) nur auf sehr speziellen/ wenigen Aufgaben getestet. Ich will für mehrere Aufgaben gucken, ob sich CookBERT für conversational agents im Kochbereich eignet, indem für mehrere relevante Tasks ausgewertet wird.

Research question:

* **“What are the effects of domain adaptation in the performance of a pre-trained German BERT model on German legal downstream tasks?”. – Effects of inserting domain vocabulary …**
* ***How does cooking domain adaption of BERT affect the performance of CA-relevant tasks in/ from this domain?***

Frummet

* In situ-study um die Informationsbedürfnisse der Nutzer beim Kochen zu untersuchen
* „our results provide an in-depth understanding oft he information needs occuring in the domain of home cooking“
  + They provide a detailed hierarchical taxonomy that shows a variety of different needs that can occur
  + Im Rahmend der Arbeit ist auch der CookversationalDatensatz entsanden, welcher in section … besprochen wird.
  + They also try to automatically identify those information needs based on their dataset. 🡪 BERT models performed best but still not good enough 🡪 sie meinen, dass das einbeziehen von Kontext hilfreich ist
* Im Paper wurden zudem noch unterschiedliche Machine Learning Ansätze und deren Performance bei der Klassifikation der Informationsbedürfnisse auf level 1 Ebene (enthält 11 Klassen) und für unterschiedliche Conditionen (utterance without context, utterance with 1 prev turn, utterance with all prev turns) verglichen.
* They found that the GermanBERT model (cite) performed best among the other approaches
* When looking at the conditions, where the utterance was prepended with 1 or all previous utterances, no significant differences between the models performance was found.
* During the coding process, they found context to be an important factor for identifying information needs and thus tested for the three conditions
* They applied multiple baseline () and BERT-based models (GermanBERT and two multilingual BERT models)
* Similar to our results, Ren et al. [68] and Aliannejadi et al. [5] showed that including more context, i.e. conversational history, improves the classification performance of the current turn.
* Including the context in the form of previous turns, however, significantly improved results in competence-oriented needs Cooking technique and Preparation.
* Beides Seite 21
* Auch wenn es beim best-performendsten Model (GermanBERT, welches dann für die 3 conditions applied wurde) overall keine signifikanten unterschiede bei der Performance für die 3 conditions gibt, kann das hinzufügen von kontext in form von previous turns trotzdem für bestimmte information needs signifikant die Performance verbessern. Und wird deshalb als sinnvoll erachtet
* Das aber alles eher schon vorher in related work packen und nicht hier beim datensatz
* **Der Vorschlag, dass BERT Domänenspezifisches Wissen fehlt, wurde auch von Schwabl gemacht!**

Sabrina paper

* Untersucht, wie Nutzer mit conversational agent for recipe recommendation interagieren
* Wizard of oz Methode wurde verwendet
  + 28 participants für zwei conditions: audio based und text based
  + The participants interacted with the wizard (they thought it was a bot) to complete three recipe finding tasks
* Propose Framework for the conversationa flow in a future possible conversational agent for recipe recommendation
* Their results are mostly encouraging for future development of such a conversational agent, but also provide insights into the complexity of building such a system
  + Many conversations oft he study followed this framework