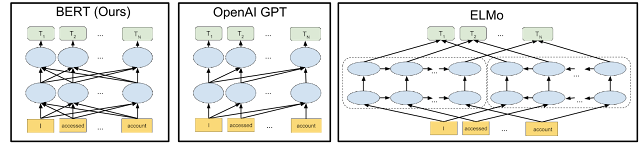
# BERT

* Published 2018 by google AI, led to a small revolution
* Huge neural network based on transformer architecture and pretrained on 3.3 Billion words
* Builds upon previous approaches on pretraining contextual representations: ELMO, GPT; Semi-supervised sequence learning, ULMFit

### Pretrained contextual representations

* + One of the biggest challenges in NLP is shortage of training data. Most task-specific datasets contain only a few hundred or thousand human-labeled training examples
  + NN brauchen aber teilweise riesige Datenmenge, um etwas gutes zu lernen
  + Um diese Lücke zu schließen, haben researcher eine Reihe an Approaches entwickelt, um general purpose language representation models mithilfe der riesigen Menge an unannotierten Text im Web, vorzutrainieren, known as pre-training.
  + Diese vortrainierten Models können dann fine-tuned werden auf small-data NLP tasks like question answering or sentiment analysis, resulting in substantial accuracy improvements compared to training these datasets from scratch
  + Pre-trained representations können entweder context-free oder contextual sein. Contextual können wiederum unidirectional or bidirectional sein.
  + Context-free: Word2Vec or GloVe erzeugen word embeddings for each word in the vocab. “bank” (Sparkasse) und “bank” (Sitzbank) hätten hier dieselbe representation, unabhängig von ihrem Kontext
  + Contextual Models: generate a representation of each word based on the context of the text/ based on the other words in the sentence
  + Unidirectional Model: “I accessed the bank account” repräsentiert bank anhand des Kontextes “I accessed the”, aber NICHT “account”
  + Bidirectional Model: BERT; represents “bank” using both, its previous and next context – making it deeply bidirectional



* BERT ist deeply bidirectional, OpenAI GPT is unidirectional and ELMO is shallowly bidirectional (Google Blogpost2) 🡪 auch in offiziellem BERT Paper

Pretraining Objectives:

* + If bidirectionality is so powerful, why hasn’t it been done before?
  + Unidirectional models are trained by predicting each word conditioned on the previous words in a sentence (classic language modelling). Ein bidirectionales Model kann allerdings nicht auf diese Art und weise trainiert werden, since it would allow the word that’s being predicted to indirectly see itself in a multi-layer model. Deshalb werden folgende Objectives verwendet

MLM

* + Helps it learn the context in a sentence
  + Maskieren von 15% der Tokens einer Input sequenz and then condition each word bidirectionally to predict the masked words.
  + Example: “The man went to the [MASK]1 . He bougt a [MASK]2 of milk. 🡪 MASK1 = store, MASK2 = gallon
  + In der LIteratur schon länger bekannt unter dem namen “Cloze task” (cloze paper)
  + Sfasd

NSP

* + Helps it learn the relationship between two sentences, nachdem das nicht directly von dem MLM gecaptured wird.
  + Für Satz A und B muss predicted werden, ob B der tatsächlich nächste Satz von A ist (50% ja, 50% einfach ein random Satz aus dem Korpus)
* Ermöglicht Transfer-Learning

### Transfer Learning:

* + The technique of transferring knowledge gained from performing one task to another, similar one
  + Beneficial compared to the resource intensive process of training networks from scratch

### Architecture

* Based on transformer architecture
* BERT decomposes input sentence(s) into WordPiece tokens (Wu et al., 2016) 🡪 wordpiece tokenization helps improve the representation of the input vocab and reduce its size by segmenting complex words into subwords. It therefore also tackles the OOV Problem by segmenting unknown vocabs in smaller units.
* Fundamentally, BERT is a stack of Transformer encoder layers (Vaswani et al., 2017) that consist of multiple self-attention ‘‘heads’’. For every input token in a sequence, each head computes key, value, and query vectors, used to create a weighted representation. The outputs of all heads in the same layer are combined and run through a fully connected layer. Each layer is wrapped with a skip connection and followed by layer normalization.
* The conventional workflow for BERT consists of two stages: pre-training and fine-tuning. Pretraining uses two self-supervised tasks: masked languagemodeling(MLM, prediction ofrandomly masked input tokens) and next sentence prediction (NSP, predicting if two input sentences are adjacent to each other). In fine-tuning for downstream applications, one or more fully connected layers are typically added on top of the final encoder layer.
* The input embedding layers (token, position, and segment) are combined to obtain a fixed-length vector. Special token [CLS] is used for classification predictions, and [SEP] separates input segments.
* Google1 and HuggingFace (Wolf et al., 2020) provide many variants of BERT, including the original ‘‘base’’ and ‘‘large’’ versions. They vary in the number of heads, layers, and hidden state size.
* Von A Primer BERTology

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, ….)

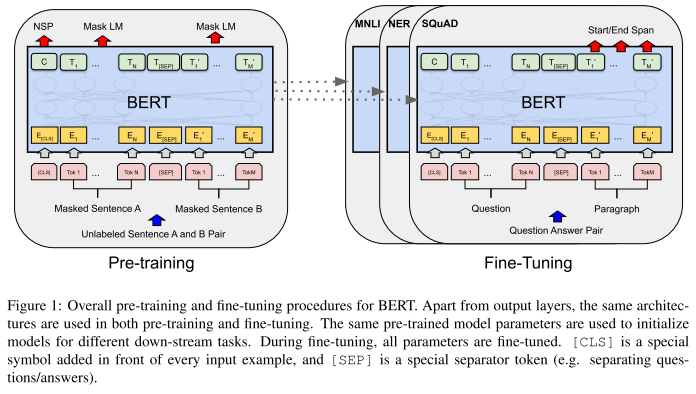


Figure : aus offiziellem BERT Paper

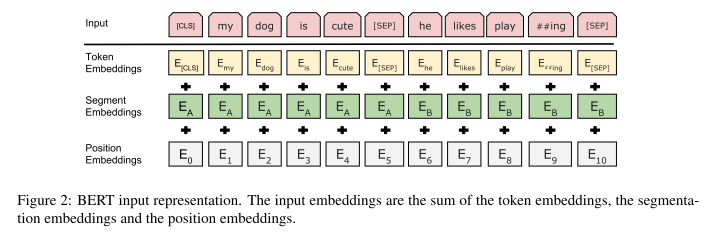


Figure : aus offiziellem BERT paper

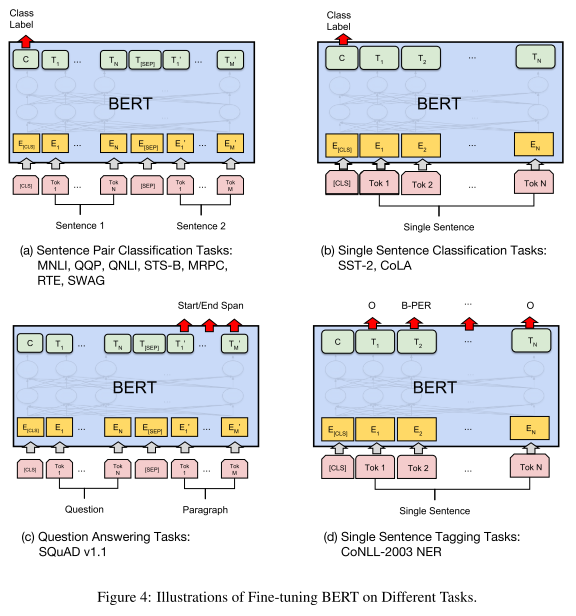


Figure : aus offiziellem BERT paper

# BERT and conversational agents

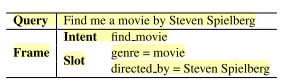
* Aufgrund der promising results of BERT for many downstream tasks, wurde es dann auch für verschiedene Aufgaben von conversational agents eingesetzt
* NLU tasks often suffer from small-scale human-labeled training data, resulting in poor generalization capability, especially for rare words. 🡪 BERT behebt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

### BERT-DST

* 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

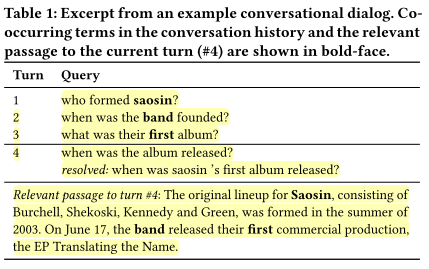
* BERT für dialogue breakdown detection
* The difficulty of developing chat-oriented dialogue systems is that such systems are required to respond to a very wide range of topics expressed by user utterances. Since it is still difficult for the current dialogue systems to continue outputting appropriate 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.