Lightweight Transformers for Conversational AI

Daniel Pressel
Interactions, LLC
dpressel@interactions.com

Wenshuo Liu
Twitter*
wenshuol@twitter.com

Michael Johnston
Alexa AI, Amazon*
mjohnstn@amazon.com

Minhua Chen
Interactions, LLC
mchen@interactions.com

Abstract

To understand how training on conversational language impacts performance of pre-trained models on downstream dialogue tasks, we build compact Transformer-based Language Models from scratch on several large corpora of conversational data. We compare the performance and characteristics of these models against BERT and other strong baselines on dialogue probing tasks. Commercial dialogue systems typically require a small footprint and fast execution time, but recent trends are in the other direction, with an ever-increasing number of parameters, resulting in difficulties in model deployment. We focus instead on training fast, lightweight models that excel at natural language understanding (NLU) and can replace existing lower-capacity conversational AI models with similar size and speed. In the process, we develop a simple but unique curriculum-based approach that moves from general-purpose to dialogue-targeted both in terms of data and objective. Our resultant models have around 1/3 the number of parameters of BERT-base and produce better representations for a wide array of intent detection datasets using linear and Mutual-Information probing techniques. Additionally, the models can be easily fine-tuned on a single consumer GPU card and deployed in near real-time production environments.

1 Introduction

The development of the Transformer (Vaswani et al., 2017) – a multi-headed attention architecture with high capacity – caused a breakthrough in the pre-training of contextualized representations for text (Radford et al., 2018; Devlin et al., 2019). This architecture can internalize large amounts of information from massive datasets, yielding powerful encoders that can be fine-tuned for various NLP tasks. The generative pre-training (GPT) model (Radford et al., 2018) used a standard language modeling objective, learning to predict the next word in a sequence, and demonstrated the ability of Transformers to learn long distance dependencies – a limitation of previous architectures. BERT (Devlin et al., 2019) later introduced a Masked Language Model (MLM) objective, where a portion of the text is masked out or perturbed, and the model learns to reconstruct those portions, yielding bi-directional representations.

Various datasets have been explored for pre-training including the Toronto BookCorpus (Zhu et al., 2015) and Wikipedia. More recently, much larger datasets have been used including Common Crawl datasets for RoBERTa (Liu et al., 2019b), T5 (Raffel et al., 2020) and others. Larger datasets facilitate more parameters (Kaplan et al., 2020; Raffel et al., 2020) and with so much available data, many recent models range from tens to hundreds of billions of parameters. These large language models (LLMs) exhibit remarkable capabilities for many tasks, but they are massive, difficult to control, and expensive to deploy. [1]

LLMs have yielded improvements over their predecessors, with little architectural modification, primarily by using larger datasets, more capacity, and training longer with more compute. The trend in the literature seems clear – use a denoising or language model objective and train on larger datasets with longer contexts and more capacity to improve NLP performance.

The main approach seems to be "more is better" without much qualification to the type of textual content that is applied. There have been some attempts to qualify the corpora used for pre-training (Mitchell et al., 2019), but attempting to limit what goes into training becomes increasingly difficult as the datasets get larger. For this work, we attempt to target conversational AI, taking into account both

* Work done while the authors were at Interactions

[1] While denoising auto-encoders such as T5 and BERT are not considered proper LMs, they are often lumped under that nomenclature, which we retain for consistency.

Proceedings of NAACL-HLT 2022: Industry Track Papers, pages 221 - 229
July 10-15, 2022 ©2022 Association for Computational Linguistics

model and dataset.

It seems reasonable to assume that some carefully curated data, like Wikipedia, should be broadly useful as it presents related concepts in close proximity with reliable structure. But for our purposes, it also seems desirable to have a large amount of data that is conversational, though its unclear how this data should be structured and how it should be balanced with non-conversational sources. Knowing that task-oriented dialogue systems often have a need to understand domain-specific concepts and proper names, it also seems reasonable to assume that sources such as consumer reviews of products and services might be helpful.

In this work, we attempt to build practical, compact models that excel for conversational AI, especially NLU. We set out to build on data sources that will be particularly useful for dialogue systems, and try to incorporate common-sense architectural modifications that should improve performance on dialogue. We focus our effort on MLM models as most NLU tasks benefit from bi-directionality (Devlin et al., 2019). We also investigate a curriculum that teaches the model a grounded foundation first in generic language, followed by increasingly complex masking over conversational data.

2 Related Work

2.1 Models targeting Dialogue

Several Transformer models have been developed specifically to target dialogue, including ConveRT (Henderson et al., 2020), and later ToD-BERT (Wu et al., 2020). The former was trained from scratch on 3 years of Reddit data (Henderson et al., 2019) using a dual-encoder with a contrastive loss function to predict the second utterance in a paired dialogue turn. ToD-BERT similarly used a contrastive loss head (in conjunction with an MLM head) to predict the next turn of dialogue (again in a dialogue pair). Unlike ConveRT, ToD-BERT was adapted to a very small corpus of only task-oriented dialogue from a pre-trained BERT checkpoint. DialoGPT (Zhang et al., 2020) is a GPT2-adapted (Radford et al., 2019) model using a corpus of conversational data. Unlike our work, they are specifically targeting response generation.

Some dialogue-targeted models have attempted to solve end-to-end task-oriented dialogue with pre-trained Transformers, including SimpleToD (Hosseini-Asl et al., 2020), a model that learns the entire process of NLU, state internalization, NLG and API calls using only a prefix-LM. This model is particularly strong on MultiWoz (Budzianowski et al., 2018), a common task-oriented dialogue benchmark, but the design is ill-suited for real-world system deployment – API calls and NLG are integrated directly into the LM, making the model difficult to adapt, control, and maintain over time. We focus instead on targeting the understanding portion of a dialogue system. As a result, our models can be easily incorporated into modular dialogue management systems.

2.2 Compact models

Distillation (Hinton et al., 2015) is a common technique for creating compact Transformers (Sanh et al., 2019; Jiao et al., 2020). In basic distillation, a larger model (the “teacher") predicts the labels on a dataset and its outputs are treated as the target distribution using the Kullback-Leibler (KL) Divergence against the predictions of a smaller “student" model. During training, the student learns to make similar predictions to the teacher (Beyer et al., 2021).

While we could attempt to train a very large Transformer from scratch and distill to a smaller one in the hopes of improved performance, this would be resource-inefficient. Alternatively, we could adapt an existing off-the-shelf pre-trained model and distill it down, but we are concerned that both the prior pre-training and distillation events could create biases on our models that would limit our ability to understand the impacts of data choices. We decided instead to training compact models from scratch without distillation.

3 Model Description

Our model is an encoder-only Transformer, similar to BERT or RoBERTa, trained primarily on full-context conversational input. Unlike most previous MLMs, our model is trained with relative attention (Shaw et al., 2018). In this approach, relative positional representations are not conditioned on the global position of the token but instead use a local relative offset embedding at every layer as part of the self-attention computation, which we hypothesize makes them more suitable for dialogue applications where the global offset in a conversation is not meaningful. Additionally, we perform layer normalization before each sub-layer and we also after the last Transformer encoder layer (Chen et al., 2018).

We also experiment with a curriculum for our MLM where the initial masking follows BERT but, later in training, switches to masked turn modeling (MTM), where we mask entire turns, token-by-token.

We train eight-layer models with eight attention heads, and a hidden size of 512. We use Byte Pair Encoding (BPE) (Sennrich et al., 2016) with a vocab containing 30,000 lower-case tokens. We also include special tokens including “[CLS]" and “[MASK]" (borrowed from BERT), “<EOS>" for end-of-sentence and “<EOU>" for end-of-utterance. We use fastBPE [2] to train, sampling 2 million posts from Reddit as the BPE corpus.

4 Datasets

In our investigation of sources of data for pre-training targeting conversational AI, we identified several potential source types. We bucket these into three basic categories:

- *foundational:* general purpose data sources, expected to provide a broad basis for pre-training
- *online reviews and customer data*: domain-informative data from single users, drawn primarily from online reviews
- *conversational*: data taken from bulletin boards and online forums capturing interactions between multiple users

We hypothesized that, used together, each data source may provide complementary information that could improve model performance and robustness for a range of dialogue tasks.

4.1 Foundational Data Sources

Dialogue systems, particularly task-oriented ones, are often required to recognize entities, their relations to one another, and to user intent. Wikipedia seems like an especially useful dataset as it makes explicit the relationship between many objects in the world. Words like “Camaro", a type of vehicle manufactured by “Chevrolet", itself a company owned by “General Motors", will all be mentioned in close proximity along with other types of vehicles, and possibly competitors. Thus we get access to a large number of proper-noun concepts and their relationships in the universe. A data source lacking encyclopedic knowledge would be unlikely to be present these primary relationships explicitly (or so thoroughly) across so many domains. Still there are several potential problems with a corpus like Wikipedia. First, the data is written formally in a manner that would rarely be encountered in actual conversations between humans. Second, queries to DBpedia to find concepts in Wikipedia show that certain domains may have quite different coverage from other domains, possibly resulting in inconsistent performance or coverage on a downstream target domain. Third (and this is a problem for any online corpus), the knowledge represented is a snapshot in time. New concepts are introduced often, and old ones may become irrelevant to daily life.

The authors of T5 created a new, large dataset from Common Crawl with broad coverage, which they refer to as “C4" (Raffel et al., 2020). The content of this corpus was subsequently analyzed and documented in (Dodge et al., 2021). Their analysis reveals that the largest sources of data within the cleaned C4 corpus are patents, Wikipedia and news sites. Based on this knowledge, a model trained on C4 would also be expected to have good coverage of concepts and their relations and strong downstream performance on problems with formal language and syntactic structure.

4.2 Online Reviews and Customer Data

We considered online reviews as a potential source of data to acquire world-knowledge useful for targeting specific domains, particularly considering entities and their relationships. For instance, for a food service application, restaurant review sites might provide some background knowledge of food items, their component parts, and in-domain co-reference knowledge. For hospitality applications, we may want to include information telling us about the properties of hotels, bed and breakfasts and resorts, and how they relate to concepts such as location, cleanliness and desirability for consumers.

4.3 Conversational Data Sources

We wished to target pre-training sources considering size and similarity to the content of our target data, including formality and structure of the lexical content. For this work, we decided to focus on English data sources only, as most of the previously available collected corpora and downstream datasets are in English.

[2] https://github.com/glample/fastBPE

We hypothesized that online forums, where users are looking for guidance regarding a product or service, would be closely related to typical task-oriented dialogue problems, and of generally high value. Also, as threads on forums can be updated over years, the lengths of forum conversations can be quite long, which may allow our models to learn rich long-distance relationships. However, even with a large number of forums, the total amount of data collected is fairly small in comparison to other online data sources. We also considered lower-quality sources of conversational data which, though dissimilar from task-oriented dialogue, might capture common discourse aspects over a large number of full conversations. We considered Twitter threads as a possible source, but the conversations tend to be very short and unfocused, and are not easy to capture using the streaming API. On the other hand, recent work has been published on Reddit as a data source (Al-Rfou et al., 2016; Henderson et al., 2019), and the scripts for obtaining this data were previously released (Henderson et al., 2019) and are reproducible with minimal cost. The corpus is quite large, and full conversations are available.

We note that, for all of our conversational data sources, author handles may be mentioned by other users, but no metadata regarding authors, locations nor any other profile information is included in the text corpus. All data collection was limited to the visible textual content of a post.

As we are particularly interested in task-oriented dialogue, we also looked into large available Wizard-of-Oz (WoZ) collected data, but found a limited selection to be available.

For all conversational data sources, we pass entire conversations into the pre-training and we use an end-of-turn marker (<EOU>) to mark new posts within a thread. In most cases its not possible (or probably even useful) to further disentangle the authors.

4.4 Combinations of data

Conversational data sources seem the most directly related to our dialogue pre-training goals. If we only considered that data, we could always provide conversational turn demarcation, and provide full conversations for each sample. We could also focus on objectives that specifically target dialogue data. However, with the non-dialogue data sources, this is not possible. As a result, its not clear how to combine the approaches effectively, or what mixture of the data we should use to support downstream applications.

We were interested in isolating the contributions of the different types of data. We trained two sets of models with slightly different datasets (version 1 and version 2). The version 1 model was trained and used internally for several months before we began work on the version 2 approach.

4.4.1 Version 1: RWD Corpus

For our first pre-training experiments, we used the full Reddit corpus from (Henderson et al., 2019), as well as 2.5 million online threads from publicly available forums, 8.2 million online reviews for restaurants and hotels, and a small amount of task-oriented dialogue (about 160,000 conversations). We determined from early experiments that Wikipedia complemented the Reddit dataset, providing better downstream fine-tuning results, so we also incorporated all of English Wikipedia. We call the resultant corpus “Reddit-Wiki-Dialogues" (RWD).

Table 1 shows a list of the datasets contained in RWD and their sizes.

4.4.2 Version 2: RF Corpus

We were interested in further isolating the impact of models trained only on conversations. While in version 1, models were trained on RWD, which is comprised of various types of data, for version 2 we attempt to better isolate the impact of conversation data only.

We introduce a conversation-only dataset containing Reddit and online forums, which we refer to as “Reddit-Forums" (RF). It does not include online reviews nor Wikipedia, nor does it use any task-oriented dialogue data.

We hypothesized that a model trained only on conversations, which are subjective in nature, might benefit from an initial pre-training with more coverage of broad concepts, so we also explore training in a curriculum that starts with a single epoch of C4 pre-training and continues on the RF corpus.

For version 2, we observed that forum data often contains quotes from previous posters, which are usually expressed in a markdown format like BB-Code [3]. For these quotes, which often juxtapose the post itself, we create additional tokens marking

[3] https://www.bbcode.org/reference.php

Label	Size	Content
Reddit	173GB	700M conversations
Wikipedia	20GB	2.7B tokens
Forums	14GB	2.5M threads
Yelp	3.9GB	6.6MB reviews
TripAdvisor	1.5GB	1.6M reviews
MetalWoz	19MB	37k conversations
DSTC7 (ubuntu)	54MB	105k conversations
DSTC8	18MB	15k conversations

Table 1: RWD Corpus

the beginning and end and rely on the attention mechanism of the Transformer in the same manner we do for end-of-sentence and end-of-turn markers. For version 2, to somewhat offset the loss of the review data and the task-oriented dialogue corpus, we collected another approx. 800k conversations from additional forum sources (yielding a total of approx 3.3M threads).

5 Training Details

We train each model using mead-baseline (Pressel et al., 2018) on a single v3 Tensor Processing Unit (TPU). [4]. To best utilize TPUs, we use bucketing based on full conversation lengths, scaling the number of samples for each bucket length so that the number of tokens is constant per batch. We use AdamW with a peak learning rate of 4e-4, a weight decay of 1e-3, and a linear warm-up of 10,000 steps followed by cosine decay over training to zero.

For version 1, we use a maximum context window of length 256, training for 1 million steps with context windows between 64 and 256 tokens.

For version 2, targeting the RF conversation-only corpus, we use a longer maximum context window of length 1024. For comparison purposes, in version 2, we train separate models using C4 only, C4 followed by RF, and RF only.

From early experiments, we determined that the MTM objective was too difficult to learn from scratch, so for MTM models, we train the first 80% following the masking algorithm for BERT, and we switch to MTM masking for the last 20%.

6 Experiments

We use the SentEval (Conneau and Kiela, 2018) approach to perform linear probing on a several intent detection datasets, including few-shot scenarios. We also perform Mutual Information-based clustering probing experiments.

6.1 Linear Probing Intent Detection

To assess the quality of our representations, we use the linear probing methods from SentEval applied to intent detection, and compare against BERT-base, ToD-BERT and SentenceBERT (Reimers and Gurevych, 2019) (an adaptation of BERT on Natural Language Inference data shown to improve embedding quality) [5]. Our analysis includes several commonly used datasets – Clinc150 (OOS) (Larson et al., 2019), PolyAI Banking77 (Casanueva et al., 2020), and the Heriot-Watt University dataset (Liu et al., 2019a). In addition to the original versions, we use 10-example and 30-example versions for each dataset to attempt understand the few-shot capabilities of our models. We also compare three internally-created intent detection datasets targeting automotive customer service, pizza customer service, and pizza ordering (shown in Table 2).

From our linear probing experiments, we find compelling evidence that our representations have internalized more useful information for dialogue, yielding better general-purpose representations. Both aspects of our training curriculum for RF models improve the overall results. The best model overall starts with C4 and continues pre-training on a conversational dataset (RF) using our MTM objective at the end of training. However, while pre-training with C4 does typically improve our MTM, the RF-only MTM model is still stronger than the version 1 model which contains all three data source types. All of our conversationally pre-trained models exhibit much better performance than the baselines. Interestingly, we observe that

[4] Each run takes 2-3 days, but curriculum branches are run from previous checkpoints, minimizing the total training time

[5] Each of these baselines has 12 heads, 12 layers, and 768 hidden units

our 8-layer, C4-only baseline is similar in performance to BERT and ToD-BERT. This might be due to the much larger size of the training set compared to BERT. Overall, our results clearly demonstrate the importance of conversational pre-training. We find very little difference in the performance of the basic MLM models using RWD versus C4+RF alone, but once coupled with the new MTM objective, the RF corpus shows a clear advantage over RWD for linear probing.

6.2 Mutual Information-based Clustering Intent Detection

In Mutual Information-based Clustering, utterances are clustered using K-means algorithm for various values of K. Then, the Adjusted Mutual Information (ANMI) score is computed between the predicted clustering and intent-based clusterings for the different settings of K (Wu and Xiong, 2020). In that work, the authors show the strength of ToD-BERT, primarily based on strong probing results using the MultiWoz dataset.

We apply the same method and compare against BERT, ToD-BERT and SentenceBERT across Intent Detection datasets. While BERT is a strong baseline for supervised downstream tasks, using Mutual Information-based Clustering, we find that ToD-BERT is significantly better than SentenceBERT which, in turn, produces consistently better representations than BERT. However, despite their much smaller size, our MTM models outperform the others by a large margin, including the MLM-only models trained on the same dataset (Figures 1, 2, 3).

For two of the datasets, the C4+RF models are the best, but for the HWU dataset, the RF-only model is better.

7 Deployment

To support deployment into our Intelligent Virtual Assistant (IVA) environment, we compared fine-tuning of our version 1 models against our production models, which operate on ASR N-best hypotheses and predict multiple outputs, representing intents and entities. For all fine-tuning, we trained on a single NVIDIA GTX1080ti and measured the joint accuracy on a real-world customer care application. We found the new models to be competitive, especially in few-shot environments. Using distillation to a single model from an ensem-

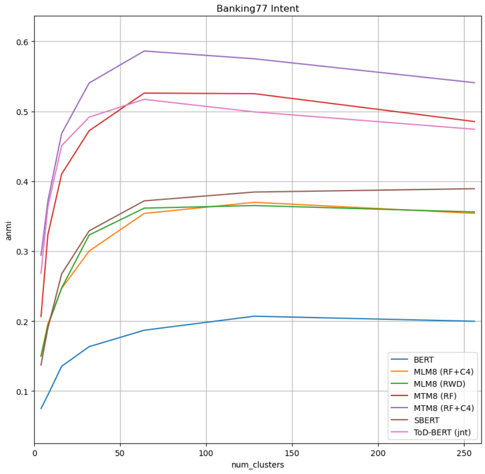

Figure 1: Banking77 Dataset ANMI

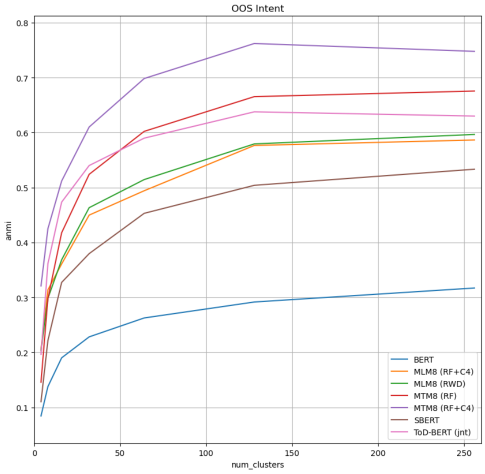

Figure 2: OOS Dataset ANMI

ID	BERT	SBERT	ToD-BERT	RWD48	C4	C4+RF	MTM RF	MTM C4+RF
AutoCS	60.54	53.79	58.98	63.45	61.06	65.52	66.25	**66.36**
PizzaCS	55.06	51.45	54.43	**60.22**	54.34	57.05	58.23	59.13
PizzaOrder	81.44	80.64	80.82	**83.39**	80.91	82.68	81.97	82.06
OOS10	72.24	80.82	79.51	84.47	77.29	84.56	85.76	**88.16**
OOS30	85.49	87.44	85.51	91.27	85.93	91.27	90.53	**91.98**
OOS	90.89	91.98	90.31	94.56	89.84	93.93	93.13	**94.93**
HWU	86.25	85.41	83.18	87.08	82.43	86.80	86.71	**89.03**
HWU10	68.31	70.54	68.96	72.58	65.61	73.42	75.74	**76.39**
HWU30	78.44	79.00	77.51	81.13	76.39	82.53	82.06	**83.36**
Bank	83.80	87.86	84.19	87.99	84.9	87.76	90.42	**90.49**
Bank10	56.14	70.71	65.42	70.42	66.82	71.56	77.27	**78.44**
Bank30	74.84	81.82	76.92	82.37	77.86	82.21	85.00	**85.84**
Avg	74.45	76.79	75.48	79.91	75.28	79.94	81.09	**82.18**

Table 2: Comparison of Model Accuracy by Probing Intent Detection Datasets

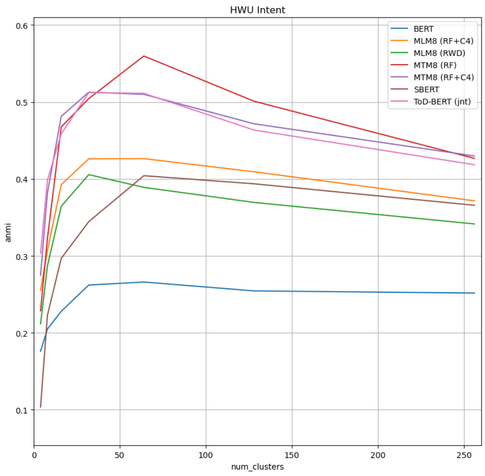

Figure 3: HWU Dataset ANMI

ble of fine-tuned MLM-based models trained on only one week of data (about 100k samples), we were able to match the performance of a production model trained on two months of data (about 650k samples). In many cases, it was also possible to truncate our models during fine-tuning, removing the top 4 layers without significant deterioration of performance. Table 3 shows the joint accuracy and speed results of our ONNX-converted models trained on a production dataset of 1.1 million noisy training samples. We observe that, even after truncating our pre-trained models to only 2 layers, they still perform better than the production system, allowing us to trade off speed and accuracy.

Model	sec/sample	Testing Acc
Production	0.0003s	83.67%
MLM 2-Layers	0.0018s	83.82%
MLM 4-layers	0.0033s	**83.94%**

Table 3: IVA Dataset Speed vs. Accuracy

8 Conclusion

Using a combination of online user conversations and sensible design choices, we are able to provide models that are compact, efficient and perform well for conversational AI. We find that conversational-only pre-training compares favorably to more traditional online data sources, but that combining the two in a curriculum can be advantageous in many cases. Additionally we find that masking whole turns later in training is also particularly helpful for learning good dialogue representations.

In future work, we will compare alternative architectures an a larger number of dialogue-oriented tasks. We are also interested in how model capacity affects downstream performance, particularly for few-shot learning. We also wish to explore the trade-offs between LM pre-training in the conversational domain versus equivalent adaptation using a contrastive loss function as used in (Vulić et al., 2021), and to further isolate the impacts of individual data sources and lexical content.

Acknowledgments

Pre-training was supported with Cloud TPUs from Google's TPU Research Cloud (TRC). TRC is an important program that empowers researchers by giving them access to compute which would otherwise be very prohibitively costly.

References

Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2016. Conversational contextual cues: The case of personalization and history for response ranking. *ArXiv*, abs/1606.00372.

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. 2021. Knowledge distillation: A good teacher is patient and consistent. *ArXiv*, abs/2106.05237.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniel Gerz, Girish Kumar, Nikola Mrksic, Georgios P. Spithourakis, Pei hao Su, Ivan Vulic, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. *ArXiv*, abs/1904.06472.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkvsi'c, Pei hao Su, Tsung-Hsien, and Ivan Vulic. 2020. Convert: Efficient and accurate conversational representations from transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2161–2174.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *ArXiv*, abs/2005.00796.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. *ArXiv*, abs/1909.10351.

Jared Kaplan, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv*, abs/2001.08361.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019a. Benchmarking natural language understanding services for building conversational agents. In *IWSDS*, pages xxx–xxx.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, page 220–229.

Daniel Pressel, Sagnik Ray Choudhury, Brian Lester, Yanjie Zhao, and Matt Barta. 2018. Baseline: A library for rapid modeling, experimentation and development of deep learning algorithms targeting nlp. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 34–40.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilyas Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI Blog*, page 1.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, page 1.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ivan Vulić, Pei-Hao Su, Samuel Coope, Daniela Gerz, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, and Tsung-Hsien Wen. 2021. ConvFiT: Conversational fine-tuning of pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1151–1168.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929.

Chien-Sheng Wu and Caiming Xiong. 2020. Probing task-oriented dialogue representation from language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5036–5051.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.