Large-scale text pre-training helps with dialogue act recognition, but not without fine-tuning

Bill Noble and **Vladislav Maraev**
Centre for Linguistic Theory and Studies in Probability
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
{bill.noble,vladislav.maraev}@gu.se

Abstract

We use dialogue act recognition (DAR) to investigate how well BERT represents utterances in dialogue, and how fine-tuning and large-scale pre-training contribute to its performance. We find that while both the standard BERT pre-training and pretraining on dialogue-like data are useful, task-specific fine-tuning is essential for good performance.

Large-scale neural language models trained on massive corpora of text data have achieved state-of-the-art results on a variety of traditional NLP tasks. Given that dialogue, especially spoken dialogue, is radically different from the kind of data these language models are pre-trained on, it is uncertain whether they would be useful for dialogue-oriented tasks. In the example from the Switchboard corpus, shown in Table 1, it is evident that the structure of dialogue is quite different from that of written text. Not only is the internal structure of contributions different—with features such as disfluencies, repair, incomplete sentences, and various vocal sounds—but the sequential structure of the discourse is different as well.

In this paper, we investigate how well one such large-scale language model, BERT (Devlin et al., 2019), represents utterances in dialogue. We use dialogue act recognition (DAR) as a proxy task, since both the internal content and the sequential structure of utterances has bearing on this task

We have two main contributions. First we find that while standard BERT pre-training is useful, the model performs poorly without fine-tuning (§3.1). Second, we find that further pre-training with data from the target domain shows promise for dialogue, but the results are mixed when pre-training with a larger corpus of dialogical data from outside the target domain (§3.2).

Speaker	DA	Utterance
A	sd	Well, I'm the kind of cook that I don't normally measure things,
A	sd	I just kind of throw them in
A	sd	and, you know, I don't to the point of, you know, measuring down to the exact amount that they say.
B	sv	That means you're a real cook.
A	bd	<Laughter> Oh, is that what it means.
A	b	Uh-huh.
A	x	<Laughter>.

Table 1: Example from the SWDA corpus (sw2827). Dialogue acts: *sd*—Statement-non-opinion, *sv*—Statement-opinion, *bd*—Downplayer, *b*—Backchannel, *x*—Non-verbal.

1 Background

1.1 Dialogue Act Recognition

The concept of a dialogue act is based on that of speech acts (Austin and Urmson, 2009). Breaking with classical semantic theory, speech act theory considers not only the propositional content of an utterance but also the actions, such as *promising* or *apologizing*, it carries out. Dialogue acts extend the concept of the speech act, with a focus on the interactional nature of most speech.

DAR is the task of labeling utterances with the dialogue act they perform from a given set of dialogue act tags. As with other sequence labeling tasks in NLP, some notion of context is helpful in DAR. One of the first performant machine learning models for DAR was a Hidden Markov Model that used various lexical and prosodic features as input (Stolcke et al., 2000). Most successful neural approaches also model some notion of context (e.g., Kalchbrenner and Blunsom, 2013; Tran et al., 2017a; Bothe et al., 2018b,a; Zhao and Kawahara, 2018).

Proceedings of the 14th International Conference on Computational Semantics, pages 166–172
June 17–18, 2021.

1.2 Transfer learning for NLP

Transfer learning techniques allow a model trained on one task—often unsupervised—to be applied to another. Since annotating natural language data is expensive, there is a lot of interest in transfer learning for natural language processing. Word vectors (e.g., Mikolov et al., 2013; Pennington et al., 2014) are a ubiquitous example of transfer learning in NLP. We note, however, that pre-trained word vectors are not always useful when applied to dialogue (Cerisara et al., 2017).

BERT, a multi-layer transformer model (Devlin et al., 2019), is pre-trained on two unsupervised tasks: *masked token prediction* and *next sentence prediction*. In masked token prediction, some percentage of words are randomly replaced with a mask token. The model is trained to predict the identity of the original token based on the context sentence. In next sentence prediction, the model is given two sentences and trained to predict whether the second sentence follows the first in the original text or if it was randomly chosen from elsewhere in the corpus. After pre-training, BERT can be applied to a supervised task by adding additional un-trained layers that take the hidden state of one or more of BERT's layers as input.

There is some previous work applying BERT to dialogue. Bao et al. (2020) and Chen et al. (2019) both use BERT for dialogue generation tasks. Similarly, Vig and Ramea (2019) find BERT useful for selecting a response from a list of candidate responses in a dialogue. Mehri et al. (2019) evaluate BERT in various dialogue tasks including DAR, and find that a model incorporating BERT outperforms a baseline model. Finally, Chakravarty et al. (2019) use BERT for dialogue act classification for a proprietary domain and achieves promising results, and Ribeiro et al. (2019) surpass the previous state-of-the-art on generic dialogue act recognition for Switchboard and MRDA corpora. This paper aims to supplement the findings of previous work by investigating how much of BERT's success for dialogue tasks is due to its extensive pre-training and how much is due to task-specific fine-tuning.

Fine-tuning vs. further in-domain pre-training We experiment with the following two transfer learning strategies (Sun et al., 2019): *further pre-training*, in which the model is trained in an un-supervised way, similar to its initial training scheme, but on data that is in-domain for the target task; and *single-task fine-tuning*, in which the model's encoder layers are optimized during training for the target task.

Whether or not the encoder model has undergone further in-domain pre-training, there remains a choice of whether to fine-tune during task training, or simply extract features from the encoder model without training it (i.e., *freezing*). Freezing the encoder model is more efficient, since the gradient of the loss function need only be computed for the task-specific layers. However, fine-tuning can lead to better performance since the encoding itself is adapted to the target task and domain.

Peters et al. (2019) investigate when it is best to fine-tune BERT for sentence classification tasks and find that when the target task is very similar to the pre-training task, fine-tuning provides less of a performance boost. We note that there is some conceptual relationship between DAR and next sentence prediction, since the dialogue act constrains (or at least is predictive of) the dialogue act that follows it. That said, the discourse strucutre of the encyclopedia and book data that makes up BERT's pre-training corpus is probably quite different from that of natural dialogue.

Switchboard	**AMI Corpus**
Dyadic	Multi-party
Casual conversation	Mock business meeting
Telephone	In-person & video
English	English
Native speakers	Native & non-native speakers
early '90s	2000s
2200 conversations 1155 in SWDA	171 meetings 139 in AMI-DA
400k utterances	118k utterances
3M tokens	1.2M tokens

Table 2: Comparison between Switchboard and the AMI Meeting Corpus

2 Data

We perform experiments on the Switchboard Dialogue Act Corpus (SWDA), which is a subset of the larger Switchboard corpus, and the dialogue act-tagged portion of the AMI Meeting Corpus (AMI-DA). SWDA is tagged with a set of 220 dialogue act tags which, following Jurafsky et al. (1997), we cluster into a smaller set of 42 tags. AMI uses a smaller tagset of 16 dialogue acts (Carletta, 2007). See Table 2 for details.

Preprocessing We make an effort to normalize transcription conventions across SWDA and AMI.

We remove disfluency annotations and slashes from the end of utterances in SWDA. In both corpora, acronyms are tokenized as individual letters. All utterances are lower-cased.

Utterances are tokenized with BERT's word piece tokenizer with a vocabulary of 30,000. To this vocabulary we added five speaker tokens and prepend each utterance with a speaker token that uniquely identifies the corresponding speaker within that dialogue.

2.1 Pre-training corpora

We also experiment with three unlabeled dialogue corpora, which we use to provide further pre-training for the BERT encoder.

The first two corpora are constructed from the same source as the dialogue act corpora. We use the SWDA portion of the un-labeled Switchboard corpus (SWBD) and the entire AMI corpus (including the 32 dialogues with no human-annotated DA tags that are not included in the DAR training set). In both cases, we exclude dialogues that are reserved for DAR testing.

We also experiment with a much larger a corpus (350M tokens) constructed from OpenSubtitles (Lison and Tiedemann, 2016). Because utterances are not labeled with speaker, we randomly assigned a speaker token to each utterance to maintain the format of the other dialogue corpora.

The pre-training corpora were prepared for the combined masked language modeling and next sentence (utterance) prediction task, as described by Devlin et al. (2019). For the smaller SWBD and AMI corpora, we generate and train on multiple epochs of data. Since there is randomness in the data preparation (e.g., which distractor sentences are chosen and which tokens are masked), we generate each training epoch separately.[1]

3 Model

We use a simple neural architecture with two components: an encoder that vectorizes utterances (BERT), and single-layer RNN sequence model that takes the utterance representations as input.[2] At each time step, the RNN takes the encoded utterance as input and its hidden state is passed to a linear layer with softmax over dialogue act tags.[3]

Conceptually, the encoded utterance represents the context-agnostic features of the utterance, and the hidden state of the RNN represents the full discourse context.

For the BERT utterance encoder, we use the $\text{BERT}_{\text{BASE}}$ model with hidden size of 768 and 12 transformer layers and self-attention heads (Devlin et al., 2019, §3.1). In our implementation, we use the un-cased model provided by Wolf et al. (2020). The RNN has a hidden layer size of 100.

3.1 Pre-training vs. fine-tuning

First, we analyze how pre-training affects BERT's performance as an utterance encoder. To do so, we consider the performance of DAR models with three different utterance encoders:

- `BERT-FT` – pre-trained + DAR fine-tuning
- `BERT-FZ` – pre-trained, frozen during DAR
- `BERT-RI` – random init. + DAR fine-tuning

`BERT-FT` is more accurate than `BERT-RI` by several percentage points on both DA corpora, suggesting that BERT's extensive pre-training does provide some useful information for DAR (Table 3). This performance boost is much more pronounced in the macro-averaged F1 score,[4] which is explained by the fact that at the tag level, pre-training has a larger impact on less frequent tags (see Figure 1 in the supplementary materials).

The `BERT-FZ` performs very poorly compared to either `BERT-FT` or `BERT-RI`, however. It is heavily biased towards the most frequent tags, which explains its especially poor macro-F1 score (Table 3). In SWDA, for example, the model with a frozen encoder predicts one of the two most common tags (Statement-non-opinion or Acknowledge) 86% of the time, whereas those two tags account for only 51% of the ground truth tags. `BERT-FT` is much less biased; it predicts the two most common tags only 59% of the time.

3.2 Impact of dialogue pre-training

Next, we assess the effect of additional dialogue pre-training on BERT's performance as an utter-

[1] For details, see the finetuning example from Hugging Face.

[2] We have experimented with LSTM as the sequence model, but the accuracy was not significantly different compared to RNN. It can be explained by the absence of longer distance dependencies on this level of our model.

[3] Other work has shown that DAR benefits from more sophisticated decoding, such as conditional random field (Chen et al., 2018) and uncertainty propagation (Tran et al., 2017b).

[4] We report both *accuracy* (which is equal to micro-averaged or class-weighted F1) and *macro-F1*, which is the unweighted average of the F1 scores of each class.

ance encoder.[5] Sun et al. (2019) has reported that performing additional pre-training on unlabeled in-domain data improves performance on classification tasks. We want to see if BERT can benefit from pre-training on dialogue data, including from data outside the immediate target domain.

For each of the target corpora (SWDA and AMI-DA), we compare four different pre-training conditions: The in-domain corpus (ID), consisting of the AMI pre-training corpus for the AMI-DA model and the SWBD pre-traning corpus for the SWDA model; the cross-domain corpus (CC), consisting of both the AMI and SWBD pre-training corpora; and finally the OpenSubtitles corpus (OS). As before, we experiment with both frozen and fine-tuned models at the task training stage.

We performed 10 epochs of pre-training on the in-domain models and 5 epochs of pre-training on the cross-domain models so that the total amount of training data was comparable. The OpenSubtitles models were trained for only one epoch but with much more total training time.

In the fine-tuned condition, additional pre-training offers a modest boost in overall accuracy and a substantial boost to the macro-F1 scores, with the cross-domain corpus providing the largest boost. In the frozen condition, only the very large OpenSubtitles corpus is helpful, suggesting that when adapting BERT to dialogue, the size of the corpus is more important than its quality or fidelity to the target domain. Still, pre-training provides nowhere near the performance improvement achieved by fine-tuning on the target task.

4 Discussion

A key aspiration of transfer learning is to expose the model to phenomena that are too infrequent to learn from labeled training data alone. We show some evidence of that here. Pre-trained BERT-FT performs better on infrequent dialogue acts than BERT-RI, suggesting it draws on the extensive pre-training to represent infrequent features of those utterances. Indeed, a simple lexical probe supports this explanation: in utterances where the pre-trained model is correct and the randomly initialized model is not, the rarest word is 1.9 times rarer on average than is typical of corpus as a whole.

[5] In-domain pre-training is sometimes referred to as *fine-tuning*, but we reserve that term for task-specific training on labeled data.

[6] Kozareva and Ravi (2019)

	SWDA		AMI-DA	
	F1	acc.	F1	acc.
BERT-FT	36.75	76.60	43.42	64.93
BERT+ID-FT	43.63	77.01	46.70	**68.88**
BERT+CC-FT	**47.78**	**77.35**	**48.86**	68.79
BERT+OS-FT	41.42	76.95	48.65	68.07
BERT-FZ	7.75	55.61	14.86	48.34
BERT+ID-FZ	6.46	52.30	14.48	48.18
BERT+CC-FZ	5.76	51.14	11.34	40.48
BERT+OS-FZ	**9.60**	**57.67**	**17.03**	**51.03**
BERT-RI	32.18	73.80	34.88	60.89
Majority class	0.78	33.56	1.88	28.27
SotA	-	83.1[6]	-	-

Table 3: Comparison of macro-F1 and accuracy with further in-domain (ID), cross-domain corpus (CC), and OpenSubtitles (OS) dialogue pre-training, for the frozen (FZ) and fine-tuned (FT) conditions. BERT-RI uses a randomly initialized utterance encoder with no pre-training but with fine-tuning.

In spite of that, the representations learned through pre-training are simply not performant without task-specific fine-tuning, suggesting that they are fundamentally lacking in information that is important for the dialogue context. We should note that this is in stark contrast to many other non-dialogical semantic tasks, where frozen BERT performs on par or *better* than the fine-tuned model (Peters et al., 2019).

By performing additional pre-training on a large dialogue-like corpus (OpenSubtitles), we were able to raise the performance of the frozen encoder by a small amount. This deserves further investigation. Bao et al. (2020) find that further pre-training BERT on a large-scale Reddit and Twitter corpus is helpful for response selection, but given the unimpressive results with subtitles, it remains an open question how well the text chat and social media domains transfer to natural dialogue.

There is also abundant room to investigate how speech-related information, such as laughter, prosody, and disfluencies can be incorporated into a DAR model that uses pre-trained features. Stolcke et al. (2000) showed, for example, that dialogue acts can have specific prosodic manifestations that can be used to improve dialogue act classification. Incorporating such information is crucial if models pre-trained on large-scale text corpora are to be adapted for use in dialogue applications.

References

John L. Austin and James O. Urmson. 2009. *How to Do Things with Words: The William James Lectures Delivered at Harvard University in 1955*, 2. ed., [repr.] edition. Harvard Univ. Press, Cambridge, Mass. OCLC: 935786421.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.

Chandrakant Bothe, Sven Magg, Cornelius Weber, and Stefan Wermter. 2018a. Conversational analysis using utterance-level attention-based bidirectional recurrent neural networks. *Proc. Interspeech 2018*, pages 996–1000.

Chandrakant Bothe, Cornelius Weber, Sven Magg, and Stefan Wermter. 2018b. A Context-based Approach for Dialogue Act Recognition using Simple Recurrent Neural Networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jean Carletta. 2007. Unleashing the killer corpus: Experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41(2):181–190.

Christophe Cerisara, Pavel Král, and Ladislav Lenc. 2017. On the effects of using word2vec representations in neural networks for dialogue act recognition. *Computer Speech & Language*, 47:175–193.

Saurabh Chakravarty, Raja Venkata Satya Phanindra Chava, and Edward A Fox. 2019. Dialog acts classification for question-answer corpora. In *ASAIL@ ICAIL*.

Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. 2019. Semantically Conditioned Dialog Response Generation via Hierarchical Disentangled Self-Attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3696–3709, Florence, Italy. Association for Computational Linguistics.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue Act Recognition via CRF-Attentive Structured Network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 225–234, New York, NY, USA. Association for Computing Machinery.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel Jurafsky, Liz Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Convolutional Neural Networks for Discourse Compositionality. In *Proceedings of the Workshop on Continuous Vector Space Models and Their Compositionality*, pages 119–126.

Zornitsa Kozareva and Sujith Ravi. 2019. ProSeqo: Projection Sequence Networks for On-Device Text Classification. 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 3894–3903, Hong Kong, China. Association for Computational Linguistics.

Pierre Lison and Jorg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, page 7.

Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi. 2019. Pretraining Methods for Dialog Context Representation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3836–3845, Florence, Italy. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *NIPS Proceedings*, page 9.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy. Association for Computational Linguistics.

Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2019. Deep dialog act recognition using multiple token, segment, and context information representations. *Journal of Artificial Intelligence Research*, 66:861–899.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339–373.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to Fine-Tune BERT for Text Classification? In *Chinese Computational Linguistics*, Lecture Notes in Computer Science, pages 194–206, Cham. Springer International Publishing.

Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. 2017a. A Hierarchical Neural Model for Learning Sequences of Dialogue Acts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 428–437, Valencia, Spain. Association for Computational Linguistics.

Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. 2017b. Preserving Distributional Information in Dialogue Act Classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2156, Copenhagen, Denmark. Association for Computational Linguistics.

Jesse Vig and Kalai Ramea. 2019. Comparison of Transfer-Learning Approaches for Response Selection in Multi-Turn Conversations. In *Proceedings of the Workshop on Dialog System Technology Challenges*, page 7, Honolulu, Hawaii.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tianyu Zhao and Tatsuya Kawahara. 2018. A unified neural architecture for joint dialog act segmentation and recognition in spoken dialog system. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–208.

Appendix A

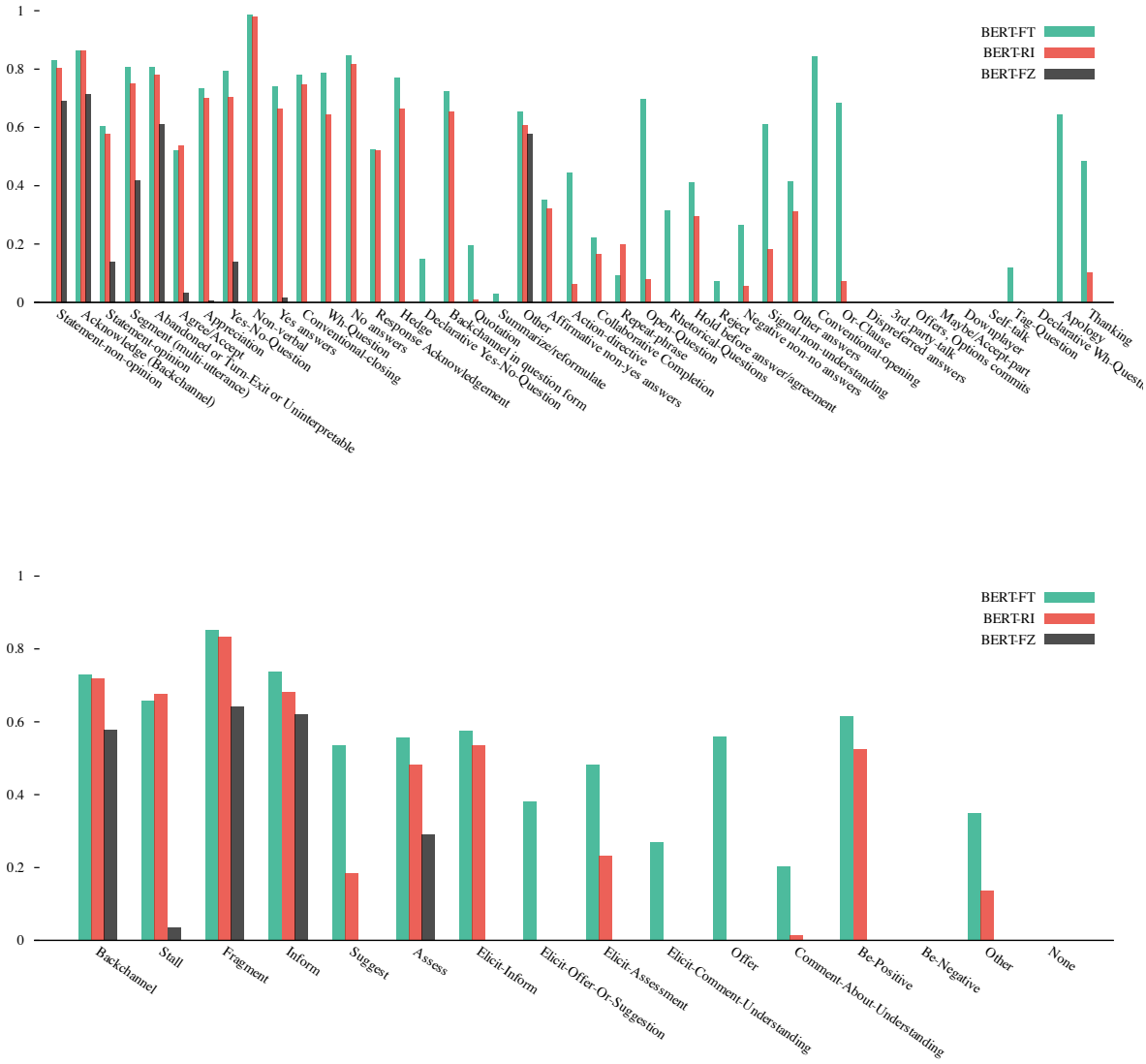

Figure 1: F1 scores by dialogue act for BERT with standard pre-training and DAR fine-tuning (BERT-FT) vs. the same model without pre-training (BERT-RI) and without fine-tuning (BERT-FZ). Dialogue acts are ordered with the most common on the left.