Example-Driven Intent Prediction with Observers

Shikib Mehri[♠*] **Mihail Eric**[♢]
[♠]Language Technologies Institute, Carnegie Mellon University, [♢]Amazon Alexa AI
amehri@cs.cmu.edu, mihaeric@amazon.com

Abstract

A key challenge of dialog systems research is to effectively and efficiently adapt to new domains. A scalable paradigm for adaptation necessitates the development of generalizable models that perform well in few-shot settings. In this paper, we focus on the intent classification problem which aims to identify user intents given utterances addressed to the dialog system. We propose two approaches for improving the generalizability of utterance classification models: (1) observers and (2) example-driven training. Prior work has shown that BERT-like models tend to attribute a significant amount of attention to the [CLS] token, which we hypothesize results in diluted representations. Observers are tokens that are not attended to, and are an alternative to the [CLS] token as a semantic representation of utterances. Example-driven training learns to classify utterances by comparing to examples, thereby using the underlying encoder as a sentence similarity model. These methods are complementary; improving the representation through observers allows the example-driven model to better measure sentence similarities. When combined, the proposed methods attain state-of-the-art results on three intent prediction datasets (BANKING77, CLINC150, HWU64) in both the full data and few-shot (10 examples per intent) settings. Furthermore, we demonstrate that the proposed approach can transfer to new intents and across datasets without any additional training.

1 Introduction

Task-oriented dialog systems aim to satisfy a user goal in the context of a specific task such as booking flights (Hemphill et al., 1990), providing transit information (Raux et al., 2005), or acting as a tour guide (Budzianowski et al., 2018). Task-oriented dialog systems must first *understand* the user's goal by extracting meaning from a natural language utterance. This problem is known as *intent prediction* and is a vital component of task-oriented dialog systems (Hemphill et al., 1990; Coucke et al., 2018). Given the vast space of potential domains, a key challenge of dialog systems research is to effectively and efficiently adapt to new domains (Rastogi et al., 2019). Rather than adapting to new domains by relying on large amounts of domain-specific data, a scalable paradigm for adaptation necessitates the development of generalizable models that perform well in few-shot settings (Casanueva et al., 2020; Mehri et al., 2020).

The task of intent prediction can be characterized as a two step process: (1) **representation** (mapping a natural language utterance to a semantically meaningful representation) and (2) **prediction** (inferring an intent given a latent representation). These two steps are complementary and interdependent, thereby necessitating that they be jointly improved. Therefore, to enhance the domain adaptation abilities of intent classification systems we propose to (1) improve the representation step through **observers** and (2) improve the prediction step through **example-driven training**.

While BERT (Devlin et al., 2018) is a strong model for natural language understanding tasks (Wang et al., 2018), prior work has found a significant amount of BERT's attention is attributed to the *[CLS]* and *[SEP]* tokens, though these special tokens do not attribute much attention to the words of the input until the last layer (Clark et al., 2019; Kovaleva et al., 2019). Motivated by the concern that attending to these tokens is causing a dilution of representations, we introduce **observers**. Rather than using the latent representation of the *[CLS]* token, we instead propose to have tokens which *attend to the words of the input* but *are not attended to*. In this manner, we disentangle BERT's attention with the objective of improving the semantic content captured by the utterance representations.

[*] Work done while Shikib was at Amazon

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2979–2992
June 6–11, 2021.

A universal goal of language encoders is that inputs with similar semantic meanings have similar latent representations (Devlin et al., 2018). To maintain consistency with this goal, we introduce **example-driven training** wherein an utterance is classified by measuring similarity to a set of examples corresponding to each intent class. While standard approaches implicitly capture the latent space to intent class mapping in the learned weights (i.e., through a classification layer), example-driven training makes the prediction step an explicit non-parametric process that reasons over a set of examples. By maintaining consistency with the universal goal of language encoders and explicitly reasoning over the examples, we demonstrate improved generalizability to unseen intents and domains.

By incorporating both observers and example-driven training on top of the CONVBERT model[1] (Mehri et al., 2020), we attain state-of-the-art results on three intent prediction datasets: BANKING77 (Casanueva et al., 2020), CLINC150 (Larson et al., 2019), and HWU64 (Liu et al., 2019) in both full data and few-shot settings. To measure the generalizability of our proposed models, we carry out experiments evaluating their ability to transfer to new intents and across datasets. By simply modifying the set of examples during evaluation and without any additional training, our example-driven approach attains strong results on both transfer to unseen intents and across datasets. This speaks to the generalizability of the approach. Further, to demonstrate that observers mitigate the problem of diluted representations, we carry out probing experiments and show that the representations produced by observers capture more semantic information than the *[CLS]* token.

The contributions of this paper are as follows: (1) we introduce observers in order to avoid the potential dilution of BERT's representations, by disentangling the attention, (2) we introduce example-driven training which explicitly reasons over a set of examples to infer the intent, (3) by combining our proposed approaches, we attain state-of-the-art results across three datasets on both full data and few-shot settings, and (4) we carry out experiments demonstrating that our proposed approach is able to effectively transfer to unseen intents and across datasets without any additional training.

[1] https://github.com/alexa/DialoGLUE/

2 Methods

In this section, we describe several methods for the task of intent prediction. We begin by describing two baseline models: a standard BERT classifier (Devlin et al., 2018) and CONVBERT with task-adaptive masked language modelling (Mehri et al., 2020). The proposed model extends the CONVBERT model of Mehri et al. (2020) through observers and example-driven training. Given the aforementioned two step characterization of intent prediction, observers aim to improve the representation step while example-driven training improves the prediction step.

2.1 BERT Baseline

Across many tasks in NLP, large-scale pre-training has resulted in significant performance gains (Wang et al., 2018; Devlin et al., 2018; Radford et al., 2018). To leverage the generalized language understanding capabilities of BERT for the task of intent prediction, we follow the standard fine-tuning paradigm. Specifically, we take an off-the-shelf BERT-base model and perform end-to-end supervised fine-tuning on the task of intent prediction.

2.2 Conversational BERT with Task-Adaptive MLM

Despite the strong language understanding capabilities exhibited by pre-trained models, modelling dialog poses challenges due to its intrinsically goal-driven, linguistically diverse, and often informal/noisy nature. To this end, recent work has proposed pre-training on open-domain *conversational data* (Henderson et al., 2019; Zhang et al., 2019b). Furthermore, task-adaptive pre-training wherein a model is trained in a self-supervised manner on a dataset prior to fine-tuning on the same dataset, has been shown to help with domain adaptation (Mehri et al., 2019; Gururangan et al., 2020; Mehri et al., 2020). Our models extend the CONVBERT model of Mehri et al. (2020) which (1) pre-trained the BERT-base model on a large open-domain dialog corpus and (2) performed task-adaptive masked language modelling (MLM) as a mechanism for adapting to specific datasets.

2.3 Observers

The pooled representation of BERT-based models is computed using the *[CLS]* token. Analysis of BERT's attention patterns has demonstrated that a significant amount of attention is attributed to

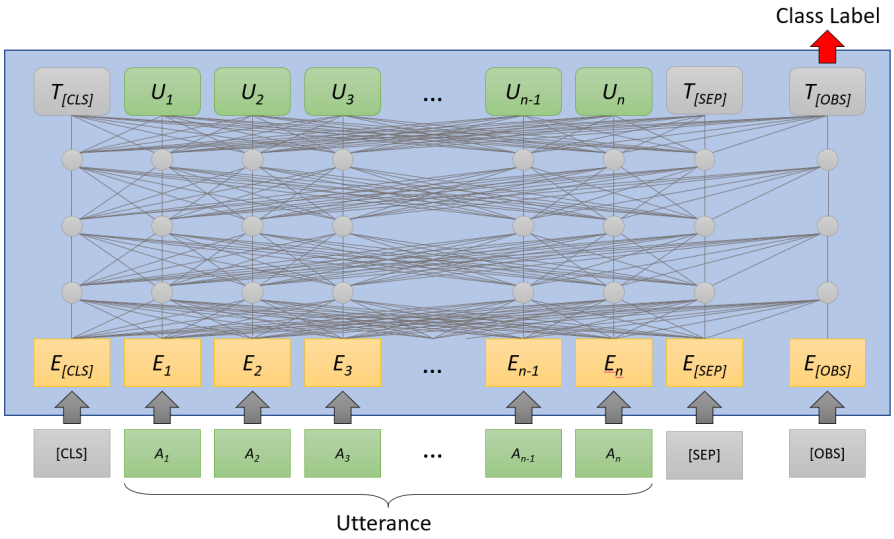

Figure 1: A visualization of the observers. The observer node attends to other tokens at each layer, however it is never attended to. While this figure only depicts one observer – we include multiple observers and average their final representation.

the *[CLS]* and *[SEP]* tokens (Clark et al., 2019; Kovaleva et al., 2019). It is often the case that over half of the total attention is to these tokens (Clark et al., 2019). Furthermore, the *[CLS]* token primarily attends to itself and *[SEP]* until the final layer (Kovaleva et al., 2019). It is possible that attending to these special BERT tokens, in combination with the residual connections of the BERT attention heads, is equivalent to a no-op operation. However, it is nonetheless a concern that this behavior of attending to tokens with no inherent meaning (since *[CLS]* does not really attend to other words until the final layer) results in the latent utterance level representations being diluted.

We posit that a contributing factor of this behavior is the entangled nature of BERT's attention: i.e., the fact that the *[CLS]* token *attends* to words of the input and is *attended to* by the words of the input. This entangled behavior may inadvertently cause the word representations to attend to *[CLS]* in order to better resemble its representation and therefore make it more likely that the *[CLS]* token attends to the word representations. In an effort to mitigate this problem and ensure the representation contains more of the semantic meaning of the utterance, we introduce an extension to traditional BERT fine-tuning called *observers*.

Observers, pictured in Figure 1, *attend to the tokens of the input utterance* at every layer of the BERT-based model however they are *never attended to*. The representation of the observers in the last layer is then used as the final utterance level representation. In this manner, we aim to disentangle the relationship between the representation of *each word* in the input and the *final utterance level representation*. By removing this bi-directional relationship, we hope to avoid the risk of diluting the representations (by inadvertently forcing them to attend to a meaningless *[CLS]* token) and therefore capture more semantic information in the final utterance level representation. Throughout our experiments we use 20 observer tokens (which are differentiated only by their position embeddings) and average their final representations. The positions of the observer tokens is consistent across all utterances (last 20 tokens in the padded sequence). Specifically, the concept of observers modifies $\mathcal{F}$ in Equations 1 and 2. While we maintain the BERT-based model architecture, we instead produce the utterance level representation by averaging the representations of the observer tokens and using that for classification rather than the *[CLS]* token.

2.4 Example-Driven Training

A universal goal of language encoders is for inputs with similar semantic meanings to have similar latent representations. BERT (Devlin et al., 2018) has been shown to effectively identify similar sentences (Reimers and Gurevych, 2019) even without additional fine-tuning (Zhang et al., 2019a). Through example-driven training, we aim to reformulate the task of intent prediction to be more consistent with

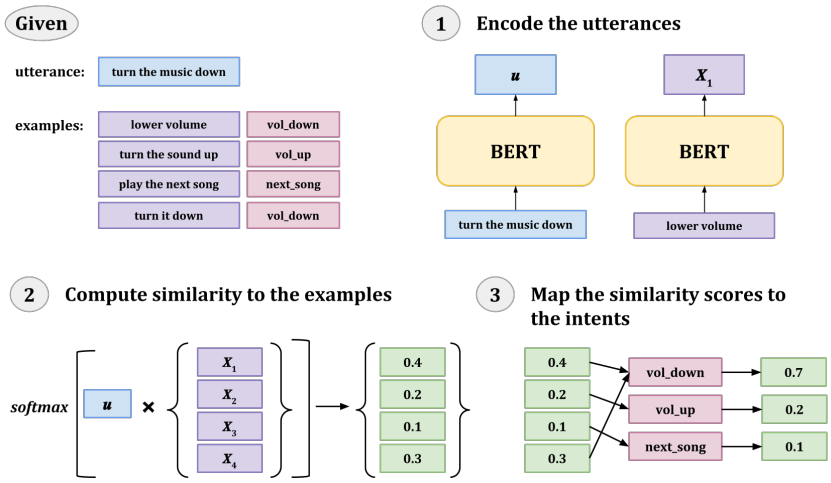

Figure 2: A visualization of the three step process of computing a probability distribution over the set of intents in our example-driven formulation.

this universal goal of language encoders.

Using a BERT-like encoder, we train an intent classification model to (1) measure the similarity of an utterance to a set of examples and (2) infer the intent of the utterance based on the similarity to the examples corresponding to each intent. Rather than implicitly capturing the *latent space* to *intent class* mapping in our learned weights (i.e., through a classification layer), we make this mapping an explicit non-parametric process that reasons over a set of examples. Our formulation, similar to metric-based meta learning (Koch et al., 2015), only performs gradient updates for the language encoder, which is trained for the task of *sentence similarity*. Through this example-formulation, we hypothesize that the model will better generalize in few-shot scenarios, as well as to rare intents.

We are given (1) a language encoder $\mathcal{F}$ that encodes an utterance to produce a latent representation, (2) a natural language utterance utt, and (3) a set of n examples $\{(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n)\}$ where $\mathrm{x}_{1,\ldots,n}$ are utterances and $y_{1,\ldots,n}$ are their corresponding intent labels. With $\mathcal{F}$ being a BERT-like model, the following equations describe example-driven intent classification:

$$\boldsymbol{u} = \mathcal{F}(utt) \quad (1)$$

$$\boldsymbol{X}_i = \mathcal{F}(\mathrm{x}_i) \quad (2)$$

$$\boldsymbol{\alpha} = \mathrm{softmax}(\boldsymbol{u}^T \cdot \boldsymbol{X}) \quad (3)$$

$$P(c) = \sum_{i:\, y_i = c} \boldsymbol{\alpha}_i \quad (4)$$

The equations above describe a non-parametric process for intent prediction. Instead, through the example-driven formulation (visualized in Figure 2), the underlying language encoder (e.g., BERT) is being trained for the task of sentence similarity. A universal goal of language encoders is that inputs with similar semantic meaning should have similar latent representations. By formulating intent prediction as a sentence similarity task, we are adapting BERT-based encoders in a way that is consistent with this universal goal. We hypothesize that in contrast to the baseline models, this formulation facilitates generalizability and has the potential to better transfer to new intents and domains.

At training time, we populate the set of examples in a two step process: (i) for each intent class that exists in the training batch, we sample one *different* utterance of the same intent class from the *training set* and (ii) we randomly sample utterances from the training set until we have a set of examples that is double the size of the training batch size (128 example utterances). During inference, our example set is comprised of all the utterances in the training data.

3 Experiments

3.1 Datasets

We evaluate our methods on three intent prediction datasets: BANKING77 (Casanueva et al., 2020), CLINC150 (Larson et al., 2019), and HWU64 (Liu et al., 2019). These datasets span several domains and consist of many different intents, making them more challenging and more reflective of commercial settings than commonly used intent predic-

tion datasets like SNIPs (Coucke et al., 2018). BANKING77 contains 13,083 utterances related to banking with 77 different fine-grained intents. CLINC150 contains 23,700 utterances spanning 10 domains (e.g., travel, kitchen/dining, utility, small talk, etc.) and 150 different intent classes. HWU64 includes 25,716 utterances for 64 intents spanning 21 domains (e.g., alarm, music, IoT, news, etc.).

Casanueva et al. (2020) forego a validation set for these datasets and instead only use a training and testing set. We instead follow the setup of Mehri et al. (2020), wherein a portion of the training set is designated as the validation set.

3.2 Experimental Setup

We evaluate in two experimental settings following prior work (Casanueva et al., 2020; Mehri et al., 2020): (1) using the full training set and (2) using 10 examples per intent or approximately 10% of the training data. In both settings, we evaluate on the validation set at the end of each epoch and perform early stopping with a patience of 20 epochs for a maximum of 100 epochs. Since the few-shot experiments are more sensitive to initialization and hyperparameters, we repeat the few-shot experiments 5 times and take an average over the experimental runs. For the few-shot settings, our models use *only* the few-shot training data for both masked language modelling and as examples at inference time in the example-driven models (i.e., they do not see any additional data). Our experiments with observers all use 20 observers, however we include an ablation in the appendix (Table 6; see supplementary materials).

3.3 Results

Our experimental results, as well as the results obtained by Casanueva et al. (2020) and Mehri et al. (2020) are shown in Table 1. Combining observers and example-driven training results in (1) SoTA results across the three datasets and (2) a significant improvement over the BERT-base model, especially in the few-shot setting (**+5.02%** on average).

Furthermore, the results show that the use of observers is particularly conducive to the example-driven training setup. Combining these two approaches gains strong improvements over the ConvBERT + MLM model (few-shot: **+4.98%**, full data: **+0.41%**). However, when we consider the two proposed approaches independently, there is no consistent improvement for both example-driven (few-shot: **-0.46%** full data: **+0.24%**) and observers (few-shot: **+0%**, full data: **-0.42%**). The fact that these two methods are particularly conductive to each other signifies the importance of using them jointly. The representation step of intent prediction is tackled by observers, which aim to better capture the semantics of an input by disentangling the attention and therefore avoiding the dilution of the representations. The prediction step, is improved through example-driven training which uses the underlying BERT-based model to predict intents by explicitly reasoning over a set of examples. This characterization highlights the importance of jointly addressing both steps of the process simultaneously. Using observers alone does not lead to significant improvements because the linear classification layer cannot effectively leverage the improved representations. Using example-driven training alone does not lead to significant improvements because the *[CLS]* representations do not capture enough of the underlying utterance semantics. The enhanced semantic representation of observers is necessary for example-driven training: by improving the latent representations of utterances, it is easier to measure similarity in the set of examples.

4 Analysis

This section describes several experiments that were carried out to show the unique benefits of observers and example-driven training, as well as to validate our hypothesis regarding the two methods. First, we show that with the example-driven formulation for intent prediction, we can attain strong performance on intents unseen during training. Next, we show that the generalization to new intents transfers across datasets. Next, we carry out a probing experiment that demonstrates that the latent representation of the observers contains greater semantic information about the input. Finally, we discuss an ablation over the number of observers used which demonstrates that the benefit of observers is primarily a consequence of the disentangled attention.

4.1 Transfer to Unseen Intents

By formulating intent prediction as a sentence similarity task, the example-driven formulation allows for the potential to predict intents that are unseen at training time. We carry out experiments in the few-shot setting for each dataset, by (1) randomly removing 4 - 10 intent classes when training in an

Model	BANKING77		CLINC150		HWU64	
	Few	**Full**	**Few**	**Full**	**Few**	**Full**
Prior Work						
USE* (Casanueva et al., 2020)	84.23	92.81	90.85	95.06	83.75	91.25
CONVERT* (Casanueva et al., 2020)	83.32	93.01	92.62	97.16	82.65	91.24
USE+CONVERT* (Casanueva et al., 2020)	85.19	93.36	93.26	97.16	85.83	92.62
BERT-BASE (Mehri et al., 2020)	79.87	93.02	89.52	95.93	81.69	89.97
CONVBERT (Mehri et al., 2020)	83.63	92.95	92.10	97.07	83.77	90.43
CONVBERT + MLM (Mehri et al., 2020)	83.99	93.44	92.75	97.11	84.52	92.38
Proposed Models						
CONVBERT + MLM + *Example*	84.09	**94.06**	92.35	97.11	83.44	92.47
CONVBERT + MLM + *Observers*	83.73	92.83	92.47	96.76	85.06	92.10
CONVBERT + MLM + *Example* + *Observers*	**85.95**	93.83	**93.97**	**97.31**	**86.28**	**93.03**

Table 1: Accuracy scores ($\times 100\%$) on all three intent detection data sets with varying number of training examples (**Few:** 10 training utterances per intent; **Full:** full training data). The full data results of Casanueva et al. (2020) are trained on more data as they forego a validation set. We follow the setup of Mehri et al. (2020), wherein a portion of the training set is used as the validation set. Results in bold-face are statistically significant by t-test ($p < 0.01$).

Model	BANKING77	CLINC150	HWU64
BERT-BASE (OFF-THE-SHELF)	19.50	26.50	26.56
CONVBERT (OFF-THE-SHELF)	19.50	26.50	26.56
CONVBERT + MLM + *Example*	67.36	79.69	62.24
CONVBERT + MLM + *Example* + *Observers*	**84.87**	**94.35**	**85.32**
BEST FULLY TRAINED MODEL	85.95	93.97	86.28

Table 2: Accuracy scores ($\times 100\%$) for transferring to unseen intents averaged over 30 runs wherein 4-10 intents are removed from the few-shot setting during training and added back in during evaluation. The last row corresponds to the best results that were trained with all of the intents, shown in Table 1. Note that the non example-driven models are incapable of predicting unseen slots, and their perform is equivalent to random chance.

example-driven manner, (2) adding the removed intents back to the set of examples during evaluation and (3) reporting results only on the unseen intents. We repeat this process 30 times for each dataset and the results are reported in Table 2. It should be noted that we do not perform MLM training on the utterances corresponding to the unseen intents.

These results demonstrate that the example-driven formulation generalizes to new intents, without having to re-train the model. The performance on the unseen intents approximately matches the performance of the best model which has seen all intents (denoted BEST FULLY TRAINED MODEL in Table 2). These results highlight a valuable property of the proposed formulation: namely, that new intent classes can be added in an online manner without having to re-train the model. While the off-the-shelf BERT-base and CONVBERT models, which are not at all fine-tuned on the datasets, are able to identify similar sentences to some extent – training in an example-driven manner drastically improves performance.

The addition of observers, in combination with example-driven training, significantly improves performance on this experimental setting (**+18.42%**). This suggests that the observers generalize better to unseen intents, potentially because the observers are better able to emphasize words that are key to differentiating between intents (e.g., *turn the volume **up*** vs *turn the volume **down***).

4.2 Transfer Across Datasets

While transferring to unseen intents is a valuable property, the unseen intents in this experimental setting are still from the same domain. To further evaluate the generalizability of our models, we carry out experiments evaluating the ability of models to transfer to *other datasets*. Using the full data setting with 10 training utterances per intent, we (1) train a model on a dataset and (2) evaluate the models on a new dataset, using the training set of the new dataset as examples during inference. In this manner, we evaluate the ability of the models to transfer to unseen intents and domains without additional training.

The results in Table 3 demonstrate the ability of the the model with obsevers and example-driven training to transfer to new datasets, which consist of both unseen intents and unseen domains. These results show that the example-driven model performs reasonably well even when transferring to domains and intents that were not seen at training time. These results, in combination with the results shown in Table 2 speak to the generalizability of the proposed methods. Specifically, by formulating intent prediction as a sentence similarity task through example-driven training, we are maintaining consistency with a universal goal of language encoders (i.e., that utterances with similar semantic meanings have similar latent representations) that effectively transfers to new settings.

4.3 Observers Probing Experiment

We hypothesized that by disentangling the attention in BERT-based models, the observers would avoid the dilution of representations (which occurs because words attend to a meaningless *[CLS]* token) and therefore better capture the semantics of the input. We validate this hypothesis through the experimental evidence presented in Table 2 wherein the use of observers results in a significant performance improvement on unseen intents. To demonstrate that observers better capture the semantics of an input, we carry out a probing experiment using the *word-content* task of Conneau et al. (2018).

We generate a latent representation of each utterance using models **with** and **without** observers. We then train a classifier layer on top of the frozen representations to reproduce the words of the input. Similar to Conneau et al. (2018), we avoid using the entire vocabulary for this probing experiment and instead use only the most frequent 1000 words for each dataset. With infrequent words, there would be uncertainty about whether the performance difference is a consequence of (1) the semantic content of the representation or (2) the quality of the probing model. Since we are concerned with measuring the former, we only consider the most frequent words to mitigate the effect of latter. Table 4 shows the micro-averaged F-1 score for the task of reproducing the words in the utterance, given the different latent representations.

A latent representation that better captures the semantics of the input utterance, will be better able to reproduce the specific words of the utterance. The results in Table 4 show that the use of observers results in latent representations that better facilitate the prediction of the input words (**+1.50** or **5%** relative improvement). These results further validate the hypothesis that the use of observers results in better latent representations.

4.4 Number of Observers

To further understand the performance of the observers, we carry out an ablation study over the number of observers. The results shown in Table 6 (in the Appendix) demonstrate that while multiple observers help, even a single observer provides benefit. This suggests that the observed performance gain is a primarily a consequence of the disentangled attention rather than averaging over multiple observers. This ablation provides further evidence that the use of observers mitigates the dilution of the utterance level representations.

5 Related Work

5.1 Intent Prediction

Intent prediction is the task of converting a user's natural language utterance into one of several predefined classes, in an effort to describe the user's intent (Hemphill et al., 1990; Coucke et al., 2018). Intent prediction is a vital component of pipeline task-oriented dialog systems, since determining the goals of the user is the first step to producing an appropriate response (Raux et al., 2005; Young et al., 2013). Prior to the advent of large-scale pre-training (Devlin et al., 2018; Radford et al., 2018), approaches for intent prediction utilize task-specific architectures and training methodologies (e.g., multi-tasking, regularization strategies) that aim to better capture the semantics of the input (Bhargava et al., 2013; Hakkani-Tür et al., 2016; Gupta et al., 2018; Niu et al., 2019).

Model	BANKING77	CLINC150	HWU64
TRAINED ON BANKING77	*93.83*	91.26	83.64
TRAINED ON CLINC150	**85.84**	*97.31*	**86.25**
TRAINED ON HWU64	77.95	**92.47**	*93.03*

Table 3: Accuracy scores ($\times$100%) for transferring across datasets (in the full data setting) using the ConvBERT + MLM + Example + Observers model. The diagonal consists of results where the model was trained and evaluated on the same dataset.

Model	BANKING77	CLINC150	HWU64
CONVBERT + MLM + *Example*	34.22	31.92	19.73
CONVBERT + MLM + *Example* + *Observers*	**35.34**	**33.84**	**21.19**

Table 4: Micro-averaged F-1 scores for the task of reproducing the words of the input (using only the most frequent 1000 words) given the different latent representations.

The large-scale pre-training of BERT makes it more effective for many tasks within natural language understanding (Wang et al., 2018), including intent prediction (Chen et al., 2019a; Castellucci et al., 2019). However, recent work has demonstrated that leveraging dialog-specific pre-trained models, such as ConveRT (Henderson et al., 2019; Casanueva et al., 2020) or CONVBERT (Mehri et al., 2020) obtains better results. In this paper, we build on a strong pre-trained conversational encoder (CONVBERT) (1) by enhancing its ability to effectively capture the semantics of the input through **observers** and (2) by re-formulating the problem of intent prediction as a sentence similarity task through **example-driven training** in an effort to better leverage the strengths of language encoders and facilitate generalizability.

5.2 Observers

Analysis of BERT's attention weights shows that a significant amount of attention is attributed to special tokens, which have no inherent meaning (Clark et al., 2019; Kovaleva et al., 2019). We address this problem by disentangling BERT's attention through the use of observers. There have been several avenues of recent work that have explored disentangling the attention mechanism in Transformers. Chen et al. (2019b) explore disentangling the attention heads of a Transformer model conditioned on dialog acts to improve response generation. He et al. (2020) disentangle the attention corresponding to the words and to the position embeddings to attain performance gains across several NLP tasks. Guo et al. (2019) propose an alternative to the fully-connected attention, wherein model complexity is reduced by replacing the attention connections with a star shaped topology.

5.3 Example-Driven Training

Recent efforts in NLP have shown the effectiveness of relying on an explicit set of *nearest neighbors* to be effective for language modelling (Khandelwal et al., 2019), question answering (Kassner and Schütze, 2020) and knowledge-grounded dialog (Fan et al., 2020). However, these approaches condition on examples only during inference or in a non end-to-end manner. In contrast, we *train* the encoder to classify utterances by explicitly reasoning over a set of examples.

The core idea of example-driven training is similar to that of metric-based meta learning which has been explored in the context of image classification, wherein the objective is to learn a kernel function (which in our case is BERT) and use it to compute similarity to a support set (Koch et al., 2015; Vinyals et al., 2016; Snell et al., 2017). In addition to being the first to extend this approach to the task of intent prediction, the key difference of example-driven training is that we use a pre-trained language encoder (Mehri et al., 2020) as the underlying sentence similarity model (i.e., kernel function). Ren and Xue (2020) leverage a triplet loss for intent prediction, which ensures that their model learns similar representations for utterances with the same intent. We go beyond this, by performing end-to-end prediction in an example-driven manner.

Our non-parametric approach for intent prediction allows us to attain SoTA results and facilitate generalizability to unseen intents and across datasets.

6 Conclusion

In order to enhance the generalizability of intent prediction models, we introduce (1) observers and (2) example-driven training. We attain SoTA results on three datasets in both full data and the few shot settings. Furthermore, our proposed approach exhibits the ability to transfer to unseen intents and across datasets without any additional training, highlighting its generalizability. We carry out a probing experiment that shows the representations produced by observers to better capture the semantic information in the input.

There are several avenues for future work. (1) Observers and example-driven training can be extended beyond intent prediction to tasks like slot filling and dialog state tracking. (2) Since observers are disentangled from the attention graph, it is worth exploring whether it possible to force each of the observers to capture a *different* property of the input (i.e., intent, sentiment, domain, etc.). (3) Our mechanism for measuring sentence similarity in our example-driven formulation can be improved.

7 Ethical Considerations

Our paper presents several approaches for improving performance on the task of intent prediction in task-oriented dialogs. We believe that neither our proposed approaches nor the resulting models have cause for ethical concerns. There is limited potential for misuse. Given the domain of our data (i.e., task-oriented dialogs), failure of the models will not result in harmful consequences. Our paper relies on significant experimentation, which may have result in a higher carbon footprint, however this is unlikely to be drastically higher than the average NLP paper.

References

Aditya Bhargava, Asli Celikyilmaz, Dilek Hakkani-Tür, and Ruhi Sarikaya. 2013. Easy contextual intent prediction and slot detection. In *2013 ieee international conference on acoustics, speech and signal processing*, pages 8337–8341. IEEE.

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. *arXiv preprint arXiv:1810.00278*.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*.

Giuseppe Castellucci, Valentina Bellomaria, Andrea Favalli, and Raniero Romagnoli. 2019. Multilingual intent detection and slot filling in a joint bert-based model. *arXiv preprint arXiv:1907.02884*.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019a. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. 2019b. 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.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Angela Fan, Claire Gardent, Chloe Braud, and Antoine Bordes. 2020. Augmenting transformers with knn-based composite memory for dialogue. *arXiv preprint arXiv:2004.12744*.

Qipeng Guo, Xipeng Qiu, Pengfei Liu, Yunfan Shao, Xiangyang Xue, and Zheng Zhang. 2019. Star-transformer. *arXiv preprint arXiv:1902.09113*.

Raghav Gupta, Abhinav Rastogi, and Dilek Hakkani-Tur. 2018. An efficient approach to encoding context for spoken language understanding. *arXiv preprint arXiv:1807.00267*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Dilek Hakkani-Tür, Gökhan Tür, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Interspeech*, pages 715–719.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Ivan Vulić, et al. 2019. Convert: Efficient and accurate conversational representations from transformers. *arXiv preprint arXiv:1911.03688*.

Nora Kassner and Hinrich Schütze. 2020. Bert-knn: Adding a knn search component to pretrained language models for better qa. *arXiv preprint arXiv:2005.00766*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*.

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, Hong Kong, China. Association for Computational Linguistics.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents. *arXiv preprint arXiv:1903.05566*.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *ArXiv*, abs/2009.13570.

Shikib Mehri, Evgeniia Razumovsakaia, Tiancheng Zhao, and Maxine Eskenazi. 2019. Pretraining methods for dialog context representation learning. *arXiv preprint arXiv:1906.00414*.

Peiqing Niu, Zhongfu Chen, Meina Song, et al. 2019. A novel bi-directional interrelated model for joint intent detection and slot filling. *arXiv preprint arXiv:1907.00390*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2019. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *arXiv preprint arXiv:1909.05855*.

Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, and Maxine Eskenazi. 2005. Let's go public! taking a spoken dialog system to the real world. In *Ninth European conference on speech communication and technology*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Fuji Ren and Siyuan Xue. 2020. Intention detection based on siamese neural network with triplet loss. *IEEE Access*, 8:82242–82254.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019b. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

A Examples

Table 6 shows examples of predictions on the HWU corpus using both observers and example-driven. These examples show that semantically similar example utterances are identified, particularly when using observers. Furthermore, the examples in Table 6 show that explicitly reasoning over examples makes intent classification models more interpretable.

B Ablations

We carry out ablations over the number of observers used to train and evaluate the models. Furthermore, we vary the number of examples seen at *inference time*, as a percentage of the set of training examples. The results shown in Table 6 demonstrate that while having more observers helps, even a single observer provides benefits. This suggests that the observed performance gain (shown in Table 1) is primarily a consequence of the disentangled attention rather than averaging over multiple observers.

The ablation over the number of examples used at inference time demonstrates that the models perform reasonably well with much fewer examples (e.g., 5% is $<$1000 examples or approximately 5 per intent). The performance drop in the few-shot experiments suggests that it is important to train with more data, however the results in Table 6 demonstrate that it not necessarily important to have all of the examples at inference time.

Utterance: It is too loud. Decrease the volume
Intent: `audio-volume-down`
Model: CONVBERT + MLM + *Example*
Predicted Intent: `audio-volume-up`
Nearest Examples:
Make sound louder (`audio-volume-up`)
Your volume is too high, please repeat that lower (`audio-volume-down`)
Too loud (`audio-volume-down`)
Can you speak a little louder (`audio-volume-up`)
Model: CONVBERT + MLM + *Example* + *Observers*
Predicted Intent: `audio-volume-down`
Nearest Examples:
It's really loud can you please turn the music down (`audio-volume-down`)
Up the volume the sound is too low (`audio-volume-up`)
Too loud (`audio-volume-down`)
Decrease the volume to ten (`audio-volume-down`)
Utterance: Please tell me about the historic facts about India
Intent: `qa-factoid`
Model: CONVBERT + MLM + *Example*
Predicted Intent: `general-quirky`
Nearest Examples:
How has your life been changed by me (`general-quirky`)
Is country better today or ten years ago? (`general-quirky`)
What happened to Charlie Chaplin? (`general-quirky`)
How does production and population affect us? (`general-quirky`)
Model: CONVBERT + MLM + *Example* + *Observers*
Predicted Intent: `qa-factoid`
Nearest Examples:
Tell me about Alexander the Great (`qa-factoid`)
Give me a geographic fact about Vilnius (`qa-factoid`)
Tell me about Donald Trump (`qa-factoid`)
I want to know more about the upcoming commonwealth games (`qa-factoid`)

Table 5: Examples of predictions on the HWU corpus with both observers and example-driven training.

Setting	BANKING77	CLINC150	HWU64
OBSERVERS = 20; EXAMPLES = 100%	**93.83**	97.31	**93.03**
OBSERVERS = 10; EXAMPLES = 100%	93.60	**97.62**	92.01
OBSERVERS = 5; EXAMPLES = 100%	93.37	97.38	92.19
OBSERVERS = 1; EXAMPLES = 100%	**93.83**	97.33	92.57
OBSERVERS = 20; EXAMPLES = 50%	**93.83**	97.31	**93.03**
OBSERVERS = 20; EXAMPLES = 10%	92.86	97.24	92.38
OBSERVERS = 20; EXAMPLES = 5%	92.82	96.95	92.57
OBSERVERS = 20; EXAMPLES = 1%	80.40	68.37	73.79

Table 6: Ablation over the number of observers (during both training and testing) and the number of examples (only during testing) used for the CONVBERT + MLM + EXAMPLE-DRIVEN + OBSERVERS model. The percentage of examples refers to the proportion of the *training set* that is used as examples for the model at evaluation time.