

Template-based Approach to Zero-shot Intent Recognition

Dmitry Lamanov¹, Pavel Burnyshev¹, Ekaterina Artemova^{1,2},
Valentin Malykh¹, Andrey Bout¹, Irina Piontkovskaya¹

¹Huawei Noah’s Ark lab, ²HSE University

Correspondence: piontkovskaya.irina@huawei.com

Abstract

The recent advances in transfer learning techniques and pre-training of large contextualized encoders foster innovation in real-life applications, including dialog assistants. Practical needs of intent recognition require effective data usage and the ability to constantly update supported intents, adopting new ones, and abandoning outdated ones. In particular, the generalized zero-shot paradigm, in which the model is trained on the seen intents and tested on both seen and unseen intents, is taking on new importance. In this paper, we explore the generalized zero-shot setup for intent recognition. Following best practices for zero-shot text classification, we treat the task with a sentence pair modeling approach. We outperform previous state-of-the-art f1-measure by up to 16% for unseen intents, using intent labels and user utterances and without accessing external sources (such as knowledge bases). Further enhancement includes lexicalization of intent labels, which improves performance by up to 7%. By using task transferring from other sentence pair tasks, such as Natural Language Inference, we gain additional improvements.

1 Introduction

User intent recognition is one of the key components of dialog assistants. With the advent of deep learning models, deep classifiers have been used throughout to recognize user intents. A common setup for the task (Chen et al., 2019; Wu et al., 2020; Casanueva et al., 2020) involves an omnipresent pre-trained language model (Devlin et al., 2018; Liu et al., 2019b; Sanh et al., 2019), equipped with a classification head, learned to predict intents. However, if the dialog assistant is extended with new skills or applications, new intents may appear. In this case, the intent recognition model needs to be re-trained. In turn, re-training the model requires annotated data, the scope of which is inherently limited. Hence, handling unseen events

defies the common setup and poses new challenges. To this end, **generalized zero-shot (GZS)** learning scenario (Xian et al., 2018), in which the model is presented at the training phase with *seen* intents and at the inference phase with both *seen* and *unseen* intents, becomes more compelling and relevant for real-life setups. The main challenge lies in developing a model capable of processing *seen* and *unseen* intents at comparable performance levels.

Recent frameworks for GZS intent recognition are designed as complex multi-stage pipelines, which involve: detecting unseen intents (Yan et al., 2020), learning intent prototypes (Si et al., 2021), leveraging common sense knowledge graphs (Siddique et al., 2021). Such architecture choices may appear untrustworthy: using learnable unseen detectors leads to cascading failures; relying on external knowledge makes the framework hardly adjustable to low-resource domains and languages. Finally, interactions between different framework’s components may be not transparent, so it becomes difficult to trace back the prediction and guarantee the interpretability of results.

At the same time, recent works in the general domain GZL classification are centered on the newly established approach of Yin et al. (2019), who formulate the task as a textual entailment problem. The class’s description is treated as a hypothesis and the text – as a premise. The GZL classification becomes a binary problem: to predict whether the hypothesis entails the premise or not. Entailment-based approaches have been successfully used for information extraction (Haneczok et al., 2021; Lyu et al.; Sainz and Rigau, 2021) and for dataless classification (Ma et al., 2021). However, the entailment-based setup has not been properly explored for GZS intent recognition to the best of our knowledge.

This paper aims to fill in the gap and extensively evaluate entailment-based approaches for GZS intent recognition. Given a meaningful intent label,

such as `reset_settings`, and an input utterance, such as *I want my original settings back*, the classifier is trained to predict if the utterance should be assigned with the presented intent or not. To this end, we make use of pre-trained language models, which encode a two-fold input (intent label and an utterance) simultaneously and fuse it at intermediate layers with the help of the attention mechanism.

We adopt three dialog datasets for GZS intent recognition and show that sentence pair modeling outperforms competing approaches and establishes new state-of-the-art results. Next, we implement multiple techniques, yielding an even higher increase in performance. Noticing that in all datasets considered, most intent labels are either noun or verb phrases, we implement a small set of lexicalizing templates that turn intent labels into plausible sentences. For example, an intent label `reset_settings` is re-written as *The user wants to reset settings*. Such lexicalized intent labels appear less surprising to the language model than intact intent labels. Hence, lexicalization of intent labels helps the language model to learn correlations between inputs efficiently. Other improvements are based on standard engineering techniques, such as hard example mining and task transferring.

Last but not least, we explore two setups in which even less data is provided by restricting access to various parts of annotated data. First, if absolutely no data is available, we explore strategies for transferring from models pre-trained with natural language inference data. Second, in the dataless setup only seen intent labels are granted and there are no annotated utterances, we seek to generate synthetic data from them by using off-the-shelf models for paraphrasing. We show that the sentence pair modeling approach to GZS intent recognition delivers adequate results, even when trained with synthetic utterances, but fails to transfer from other datasets.

The key contributions of the paper are as follows:

1. we discover that sentence pair modeling approach to GZS intent recognition establishes new state-of-the-art results;
2. we show that lexicalization of intent labels yields further significant improvements;
3. we use task transferring, training in dataless regime and conduct error analysis to investigate the strengths and weaknesses of sentence pair modeling approach.

2 Related Work

Our work is related to two lines of research: zero-shot learning with natural language descriptions and intent recognition. We focus on adopting existing ideas for zero-shot text classification to intent recognition.

Zero-shot learning has shown tremendous progress in NLP in recent years. The scope of the tasks, studied in GZS setup, ranges from text classification (Yin et al., 2019) to event extraction (Haneczok et al., 2021; Lyu et al.), named entity recognition (Li et al., 2020) and entity linking (Logeswaran et al., 2019). A number of datasets for benchmarking zero-shot methods has been developed. To name a few, Yin et al. (2019) create a benchmark for general domain text classification. SGD (Rastogi et al., 2020) allows for zero-shot intent recognition.

Recent research has adopted a scope of novel approaches, utilizing natural language descriptions, aimed at zero-shot setup. Text classification can be treated in form of a textual entailment problem (Yin et al., 2019), in which the model learns to match features from class’ description and text, relying on early fusion between inputs inside the attention mechanism. The model can be fine-tuned solely of the task’s data or utilize pre-training with textual entailment and natural language inference (Sainz and Rigau, 2021). However, dataless classification with the help of models, pre-trained for textual entailment only appears problematic due to models’ high variance and instability (Ma et al., 2021). This justifies the rising need for learnable domain transferring (Yin et al., 2020) and self-training (Ye et al., 2020), aimed at leveraging unlabeled data and alleviating domain shift between seen and unseen classes.

Intent recognition Supervised intent recognition requires training a classifier with a softmax layer on top. Off-the-shelf pre-trained language models or sentence encoders are used to embed an input utterance, fed further to the classifier (Casanueva et al., 2020). Augmentation techniques help to increase the amount of training data and increase performance (Xia et al., 2020). Practical needs require the classifier to support emerging intents. Re-training a traditional classifier may turn out resource-greedy and costly. This motivates work in (generalized) zero-shot intent recognition, i.e. handling seen and unseen intents simultaneously. Early approaches to GZS intent recog-

nition adopted **capsule networks** to learn low-dimensional representations of intents. IntentCapsNet (Xia et al., 2018) is built upon three capsule modules, organized hierarchically: the lower module extracts semantic features from input utterances. Two upper modules execute recognition of seen and unseen intents independently from each other. ReCapsNet (Liu et al., 2019a) is built upon a transformation schema, which detects unseen events and makes predictions based on unseen intents’ similarity to the seen ones. SEG (Yan et al., 2020) utilizes **Gaussian mixture models** to learn intent representations by maximising margins between them. One of the concurrent approaches, CTIR (Si et al., 2021) (Class-Transductive Intent Representations) learns **intent representations from intent labels** to model inter-intent connections. CTIR is not a stand-alone solution but rather integrates existing models, such as BERT, CNN, or CapsNet. The framework expands the prediction space at the training stage to be able to include unseen classes, with the unseen label names serving as pseudo-utterances. The current state-of-the-art performance belongs to RIDE (Siddique et al., 2021), an intent detection model that leverages **common knowledge** from ConceptNet. RIDE captures semantic relationships between utterances and intent labels considering concepts in an utterance linked to those in an intent label.

3 Sentence pair modelling for intent recognition

3.1 Problem formulation

Let \mathcal{X} be the set of utterances, $\mathcal{S} = \{y_1, \dots, y_k\}$ be the set of seen intents and $\mathcal{U} = \{y_{k+1}, \dots, y_n\}$ be the set of unseen intents. The training data consists of annotated utterances $\{x_i, y_j\}$. At the test time, the model is presented with a new utterance. In the GZS setup the model chooses an intent from both seen and unseen $y_j \in \mathcal{S} \cup \mathcal{U}$.

3.2 Our approach

A contextualized encoder is trained to make a binary prediction: whether the utterance x_i is assigned with the intent y_j or not. The model encodes the intent description and the utterance, concatenated by the separation token [SEP]. The representation of the [CLS] token is fed into a classification head, which makes the desired prediction $P(1|y_j, x_i)$. This approach follows standard sentence pair (SP) modeling setup.

ID	Template
declarative templates	
d ₁	<i>the user wants to</i> the user wants to book a hotel
d ₂	<i>tell the user how to</i> tell the user how to book a hotel
question templates	
q ₁	<i>does the user want to</i> does the user want to book a hotel
q ₂	<i>how do I</i> how do I book a hotel

Table 1: Lexicalization templates, applied to intent labels. Examples are provided for the intent label “book hotel”.

Given an intent y_j , the model is trained to make a positive prediction for an in-class utterance x_i^+ and a negative prediction for an out-of-class utterance x_i^- , sampled from another intent. At the train time, the model is trained with seen intents only $y_j \in \mathcal{S}$.

On the test time, given an utterance x_i^{test} , we loop over all intents $y_j \in \mathcal{S} \cup \mathcal{U}$ and record the probability of the positive class. Finally, we assign to the utterance x_i^{test} such y^* , that provides with the maximum probability of the positive class:

$$y^* = \arg \max_{y_j \in \mathcal{S} \cup \mathcal{U}} P(1|y_j, x_i^{test})$$

Contextualized encoders. We use RoBERTa_{base} (Liu et al., 2019b) as the main and default contextualized encoder in our experiments, as it shows superior performance to BERT (Devlin et al., 2018) in many downstream applications. RoBERTa’s distilled version, DistilRoBERTa (Sanh et al., 2019) is used to evaluate lighter, less computationally expensive models. Also, we use a pre-trained task-oriented dialogue model, TOD-BERT (Wu et al., 2020) to evaluate whether domain models should be preferred.

We used models, released by HuggingFace library (Wolf et al., 2020): roberta-base, distilroberta-base and TODBERT/TOD-BERT-JNT-V1.

Negative sampling strategies include (i) sampling negative utterances for a fixed intent, denoted as (y_j, x_i^+) , (y_j, x_i^-) ; (ii) sampling negative intents for a fixed utterance, denoted as (y_j^+, x_i) , (y_j^-, x_i) .

Both strategies support sampling with hard examples. In the first case (i), we treat an utterance

x_i^- as a hard negative one for intent y_j , if there exists such in-class utterance x_i^+ , so that the similarity between x_i^+ and x_i^- is higher than a predefined threshold. To this end, to compute semantic similarity, we make use of SentenceBERT (Reimers and Gurevych, 2019) cosine similarity. For a given positive in-class utterance, we selected the top-100 most similar negative out-of-class utterance based on the values of cosine similarity. In the second case (ii), we use the same approach to sample hard negative intents y_j^- , given an utterance x_i , assigned with the positive intent y_j^+ . Again, we compute semantic similarity between intent labels and sample an intent y_j^- with probability based on similarity score with intent y_j^+ . To justify the need to sample hard negative examples, we experiment with random sampling, choosing randomly (iii) negative utterances or (iv) negative intents.

Lexicalization of intent labels utilizes simple grammar templates to convert intent labels into natural-sounding sentences. For this aim, we utilize two types of templates: (i) declarative templates (“*the user wants to*”) and (ii) question templates (“*does the user want to*”). Most intent labels take a form of a verb phrase (VERB + NOUN⁺), such as `book_hotel` or a noun phrase (NOUN⁺), such as `flight_status`. We develop the set of rules that parses an intent label, detects whether it is a verb phrase or a noun phrase¹, and lexicalizes it using one of the templates using the following expression: *template* + VERB + *a/an* + NOUN⁺. If the intent label is recognized as a noun phrase, the VERB slot is filled with an auxiliary verb, “get”. This way, we achieve such sentences: *the user wants to book a hotel* and *does the user want to get a flight status*. The templates implemented are shown in Table 1.

Lexicalization templates were constructed from the most frequent utterance prefixes, computed for all datasets. This way, lexicalized intents sound natural and are close to the real utterances. We use declarative and question templates because the datasets consist of such utterance types. We experimented with a larger number of lexicalization templates, but as there is no significant difference in performance, we limited ourselves to two templates of each kind for the sake of brevity.

Task transferring Task transferring from other tasks to GZS intent recognition allows to estimate

¹We use a basic NLTK POS tagger to process intent labels.

whether (i) pre-trained task-specific models can be used without any additional fine-tuning, reducing the need of annotated data and (ii) pre-training on other tasks and further fine-tuning is beneficial for the final performance.

There are multiple tasks and fine-tuned contextualized encoders, which we may exploit for task transferring experiments. For the sake of time and resources, we did not fine-tune any models on our own, but rather adopted a few suitable models from HuggingFace library, which were fine-tuned on the Multi-Genre Natural Language Inference (MultiNLI) dataset (Williams et al., 2018): BERT-NLI (textattack/bert-base-uncased-MNLI), BART-NLI (bart-large-mnli), RoBERTa-NLI (textattack/roberta-base-MNLI).

Dataless classification We experiment with a dataless classification scenario, in which we train the models on synthetic data. To this end, we used three pre-trained three paraphrasing models to paraphrase lexicalized intent labels. For example, the intent label `get_alarms` is first lexicalized as *tell the user how to get alarms* and then paraphrased as *What’s the best way to get an alarm?*. Next, we merge all sentences, paraphrased with different models, into a single training set. Finally, we train the GZS model with the lexicalized intent labels and their paraphrased versions without using any annotated utterances.

The T5-based (Raffel et al., 2020) and Pegasus-based (Zhang et al., 2020) paraphraser (ramsrigouthamg/t5_paraphraser, Vamsi/T5_Paraphrase_Paws tuner007/pegasus_paraphrase) were adopted from the HuggingFace library and were used with default parameters and beam size equal to 25.

4 Datasets

SGD (Schema-Guided Dialog) (Rastogi et al., 2020) contains dialogues from 16 domains and 46 intents and provides the explicit train/dev/test split, aimed at the GZSL setup. Three domains are available only in the test set. This is the only dataset, providing short intent descriptions, which we use instead of intent labels. To pre-process the SGD dataset, we keep utterances where users express an intent, selecting utterances in one of the two cases: (i) first utterances in the dialogue and (ii) an utterance that changes the dialogue state and

Method	SGD				MultiWoZ				CLINC			
	Unseen		Seen		Unseen		Seen		Unseen		Seen	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
SEG	0.372	0.403	0.613	0.636	0.371	0.414	0.652	0.646	-	-	-	-
RIDE+PU	0.590	0.573	0.832	0.830	0.569	0.521	0.884	0.885	0.798	0.573	0.908	0.912
ZSDNN + CTIR	0.603	0.580	0.809	0.878	0.468	0.437	0.827	0.892	0.561	0.493	0.904	0.871
CapsNet + CTIR	0.567	0.507	0.897	0.912	0.481	0.404	0.903	0.906	0.530	0.572	0.866	0.883
SP RoBERTa (ours)	0.698	0.732	0.917	0.925	0.606	0.686	0.903	0.919	0.661	0.742	0.946	0.954
SP RoBERTa + templates (ours)	0.750	0.805	0.931	0.934	0.624	0.722	0.941	0.948	0.692	0.766	0.927	0.931

Table 2: Comparison of different methods. SP stands for Sentence Pair modeling approach. SP RoBERTa (ours) shows consistent improvements of F1 across all datasets for seen and unseen intents. The usage of lexicalized templates improves performance.

expresses a new intent. We use pre-processed utterances from original train/dev/test sets for the GZS setup directly without any additional splitting.

MultiWoZ 2.2 (Multi-domain Wizard of Oz) (Budzianowski et al., 2018) is treated same way as SGD: we keep utterances that express an intent and we get 27.5K utterances, spanning over 11 intents from 7 different domains. We used 8 (out of 11) randomly selected intents as seen for training. 30% utterances from seen intents. All utterances implying unseen intents are used for testing. Test utterances for seen intents are sampled in a stratified way, based on their support in the original dataset.

CLINC (Larson et al., 2019) contains 23,700 utterances, of which 22,500 cover 150 in-scope intents, grouped into ten domains. We follow the standard practice to randomly select 3/4 of the in-scope intents as seen (112 out of 150) and 1/4 as unseen (38 out of 150). The random split was made the same way as for MultiWoZ.

5 Experiments

Baselines We use **SEG**², **RIDE**³, **CTIR**⁴ as baselines, as they show the up-to-date top results on the three chosen datasets. For the RIDE model, we use the base model with a Positive-Unlabeled classifier, as it gives a significant improvement on the SGD and MultiWoZ datasets. We used Zero-Shot DNN and CapsNets along with CTIR, since

²<https://github.com/fanolabs/0shot-classification>, unfortunately were unable to run the code and adopted the published results from the paper

³<https://github.com/RIDE-SIGIR/GZS>

⁴<https://github.com/PhoebusSi/CTIR>

these two encoders perform best on unseen intents (Si et al., 2021).

Evaluation metrics commonly used for the task are accuracy (**Acc**) and **F1**. The F1 values are per class averages weighted with their respective support. Following previous works, we report results on **seen** and **unseen** intents separately. Evaluation for the test set **overall** is presented in Appendix. We report averaged results along with standard deviation for ten runs of each experiment.

Results of experiments are presented in Table 2 (see Appendix for standard deviation estimation). Our approach SP RoBERTa, when used with intent labels and utterances only, shows significant improvement over the state-of-the-art on all three datasets, both on seen and unseen intents, by accuracy and F1 measures. The only exception is unseen intents of CLINC, where our approach underperforms in terms of accuracy of unseen intents recognition comparing to RIDE. At the same time, RIDE shows a lower recall score in this setup. So, our method is more stable and performs well even when the number of classes is high.

Similarly to other methods, our method recognizes seen intents better than unseen ones, reaching around 90% of accuracy and F1 on the former. Next, with the help of lexicalized intent labels our approach yields even more significant improvement for all datasets. The gap between our approach and baselines becomes wider, reaching 14% of accuracy on SGD’s unseen intents and becoming closer to perfect detection on seen intents across all datasets. The difference between our base approach SP RoBERTa and its modification, relying on intent lexicalization, exceeds 7% on unseen in-

Method	SGD		MultiWoZ		CLINC	
	Acc	F1	Acc	F1	Acc	F1
SP RoBERTa	0.687 ± 0.018	0.716 ± 0.016	0.594 ± 0.180	0.705 ± 0.157	0.639 ± 0.038	0.731 ± 0.028
SP BERT	0.668 ± 0.001	0.701 ± 0.001	0.604 ± 0.190	0.704 ± 0.162	0.613 ± 0.023	0.694 ± 0.031
SP TOD-BERT	0.658 ± 0.055	0.724 ± 0.042	0.629 ± 0.235	0.715 ± 0.241	0.625 ± 0.029	0.704 ± 0.034
SP DistilRoBERTa	0.658 ± 0.046	0.710 ± 0.022	0.603 ± 0.208	0.701 ± 0.213	0.583 ± 0.030	0.672 ± 0.029
SP RoBERTa + random IS	0.687 ± 0.018	0.716 ± 0.016	0.594 ± 0.180	0.705 ± 0.157	0.639 ± 0.038	0.731 ± 0.028
SP RoBERTa + random US	0.677 ± 0.017	0.707 ± 0.014	0.531 ± 0.218	0.632 ± 0.217	0.658 ± 0.043	0.735 ± 0.036
SP RoBERTa + hard IS	0.741 ± 0.010	0.786 ± 0.017	0.561 ± 0.177	0.680 ± 0.136	0.590 ± 0.039	0.669 ± 0.036
SP RoBERTa + hard US	0.698 ± 0.012	0.732 ± 0.019	0.606 ± 0.244	0.686 ± 0.234	0.661 ± 0.033	0.742 ± 0.028
Zero-shot RoBERTa-NLI	0.315 ± 0.000	0.382 ± 0.000	0.090 ± 0.000	0.110 ± 0.000	0.065 ± 0.000	0.068 ± 0.000
SP RoBERTa-NLI	0.748 ± 0.026	0.801 ± 0.028	0.669 ± 0.185	0.758 ± 0.151	0.700 ± 0.040	0.771 ± 0.031
SP BERT-NLI	0.693 ± 0.017	0.738 ± 0.015	0.624 ± 0.231	0.715 ± 0.197	0.614 ± 0.035	0.695 ± 0.026
SP BART-NLI	0.789 ± 0.024	0.830 ± 0.030	0.673 ± 0.174	0.753 ± 0.143	0.770 ± 0.039	0.829 ± 0.034

Table 3: Ablation study and task transferring: comparison on unseen intents. **Top**: comparison of different contextualized encoders; **middle**: comparison of negative sampling strategies of intent sampling (IS) and utterance sampling (US); **bottom**: task transferring from the MNLI dataset, using various fine-tuned models.

tents for SGD dataset and reaches 3% on MultiWoZ ones. Notably, SP RoBERTa does not overfit on seen intents and achieves a consistent increase both on unseen and seen intents compared to previous works.

Ablation study We perform ablation studies for two parts of the SP RoBERTa approach and present the results for unseen intents in Table 3. In all ablation experiments we use the SP approach with intent labels to diminish the effect of lexicalization.

First, we evaluate **the choice of the contextualized encoder**, which is at the core of our approach (see the top part of Table 3). We choose between BERT_{base}, RoBERTa_{base}, its distilled version DistilRoBERTa, and TOD-BERT. BERT_{base} provides poorer performance when compared to RoBERTa_{base}, which may be attributed to different pre-training setup. At the same time, TOD-BERT’s scores are compatible with the ones of RoBERTa on two datasets, thus diminishing the importance of domain adaptation. A higher standard deviation, achieved for the MultiWoZ dataset, makes the results less reliable. The performance of DistilRoBERTa is almost on par with its teacher, RoBERTa, indicating that our approach can be used with a less computationally expensive model almost without sacrificing quality.

Second, we experiment with the choice of **negative sampling strategy** (see the middle part of Table 3), in which we can sample either random or hard negative examples for both intents and utterances. The overall trend shows that sampling hard examples improves over random sampling (by up

to 6% of accuracy for the SGD dataset).

Choice of lexicalization templates Table 4 demonstrates the performance of SP RoBERTa with respect to the choice of lexicalization templates. Regardless of which template is used, the results achieved outperform SP RoBERTa with intent labels. The choice of lexicalization template slightly affects the performance. The gap between the best and the worst performing template across all datasets is about 2%. The only exception is q_2 , which drops the performance metrics for two datasets. In total, this indicates that our approach must use just any of the lexicalization templates, but which template exactly is chosen is not as important. What is more, there is no evidence that declarative templates should be preferred to questions or vice versa.

Further adjustments of intent lexicalization templates and their derivation from the datasets seem a part of future research. Other promising directions include using multiple lexicalized intent labels jointly to provide opportunities for off-the-shelf augmentation at the test and train times.

Task transferring results are presented in the bottom part of Table 3. First, we experiment with zero-shot task transferring, using RoBERTa-NLI to make predictions only, without any additional fine-tuning on intent recognition datasets. This experiment leads to almost random results, except for the SGD datasets, where the model reaches about 30% correct prediction.

However, models, pre-trained with MNLI and fine-tuned further for intent recognition, gain sig-

Intent description	SGD		MultiWoZ		CLINC	
	Acc	F1	Acc	F1	Acc	F1
intent labels	0.687 ± 0.018	0.716 ± 0.016	0.594 ± 0.180	0.705 ± 0.157	0.639 ± 0.038	0.731 ± 0.028
d ₁ templates	0.750 ± 0.019	0.805 ± 0.021	0.624 ± 0.231	0.722 ± 0.175	0.692 ± 0.031	0.766 ± 0.028
d ₂ templates	0.752 ± 0.003	0.804 ± 0.006	0.610 ± 0.219	0.713 ± 0.201	0.685 ± 0.035	0.756 ± 0.031
q ₁ templates	0.765 ± 0.019	0.818 ± 0.021	0.621 ± 0.208	0.727 ± 0.174	0.670 ± 0.034	0.747 ± 0.029
q ₂ templates	0.753 ± 0.026	0.807 ± 0.026	0.599 ± 0.212	0.702 ± 0.188	0.554 ± 0.054	0.620 ± 0.055

Table 4: Comparison of different lexicalization templates, improving the performance of SP RoBERTa. Metrics are reported on unseen intents only. Each row corresponds to experiments with a single lexicalization template only, isolated from the others, i.e the row “d₁ templates” uses only the d₁ form.

Train data: intent labels +	SGD				MultiWoZ				CLINC			
	Unseen		Seen		Unseen		Seen		Unseen		Seen	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
original utterances	0.687	0.716	0.916	0.922	0.594	0.705	0.903	0.912	0.639	0.731	0.894	0.903
synthetic utterances	0.666	0.688	0.746	0.778	0.615	0.642	0.621	0.713	0.580	0.613	0.608	0.654

Table 5: Dataless classification. Metrics are reported on seen and unseen intents. Fine-tuning SP-Roberta on synthetic utterances (bottom) shows moderate decline, compared to training on real utterances (top).

nificant improvement up to 7%. The improvement is even more notable in the performance of BART-NLI, which obtains the highest results, probably, because of the model’s size.

Dataless classification results are shown in Table 5. This experiment compares training on two datasets: (i) intent labels and original utterances, (ii) intent labels and synthetic utterances, achieved from paraphrasing lexicalized intent labels. In the latter case, the only available data is the set of seen intent labels, used as input to SP RoBERTa and for further paraphrasing. Surprisingly, the performance declines moderately: the metrics drop by up to 30% for seen intents and up to 10% for unseen intents. This indicates that a) the model learns more from the original data due to its higher diversity and variety; b) paraphrasing models can re-create some of the correlations from which the model learns.

The series of experiments in transfer learning and dataless classification aims at real-life scenarios in which different parts of annotated data are available. First, in zero-shot transfer learning, we do not access training datasets at all (Table 2, Zero-shot RoBERTa NLI). Second, in the dataless setup, we access only **seen** intent labels, which we utilize both as class labels and as a source to create synthetic utterances (Table 5). Thirdly, our main experiments consider both **seen** intents and utterances available (Table 2, SP RoBERTa). In the

second scenario, we were able to get good scores that are more or less close to the best-performing model. We believe efficient use of intent labels overall and to generate synthetic data, in particular, is an important direction for future research.

6 Analysis

Error analysis shows, that SP RoBERTa tends to confuse intents, which (i) are assigned with semantically similar labels or (ii) share a word. For example, an unseen intent `get_train_tickets` gets confused with the seen intent `find_trains`. Similarly, pairs of seen intents `play_media` and `play_song` or `find_home_by_area` and `search_house` are hard to distinguish.

We checked whether errors in intent recognition are caused by utterances’ surface or syntax features. Following observations hold for the SGD dataset. Utterances, which take the form of a question, are more likely to be classified correctly: 93% of questions are assigned with correct intent labels, while there is a drop for declarative utterances, of which 90% are recognized correctly. The model’s performance is not affected by the frequency of the first words in the utterance. From 11360 utterances in the test set, 4962 starts with 3-grams, which occur more than 30 times. Of these utterances, 9% are misclassified, while from the rest of utterances, which start with rarer words, 10% are misclassified.

The top-3 most frequent 3-grams at the beginning of an utterance are *I want to*, *I would like*, *I need to*.

Stress test for NLI models (Naik et al., 2018) is a typology for the standard errors of sentence pair models, from which we picked several typical errors that can be easily checked without additional human annotation. We examine whether one of the following factors leads to an erroneous prediction: (i) word overlap between an intent label and an utterance; (ii) the length of an utterance; (iii) negation or double negation in an utterance; (iv) numbers, if used in an utterance. Additionally, we measured the semantic similarity between intent labels and user utterances through the SentenceBERT cosine function to check whether it impacts performance.

Test	Correct	Incorrect
# overlapping tokens	0.94	0.63
# tokens in utterance	14.96	13.96
# digits in utterance	0.31	0.23
# neg. words in utterance	0.03	0.02
Semantic similarity	0.22	0.21

Table 6: Stress test of SP RoBERTa predictions. An utterance is more likely to be correctly predicted if it shares at least one token with the intent labels.

Table 6 displays the stress test results for one of the runs of SP RoBERTa, trained with q_1 template on the SGD dataset. This model shows reasonable performance, and its stress test results are similar to models trained with other templates. The results are averaged over the test set. An utterance gets more likely to be correctly predicted if it shares at least one token with the intent label. However, the semantic similarity between intent labels and utterances matters less and is relatively low for correct and incorrect predictions. Longer utterances or utterances, which contain digits, tend to get correctly classified more frequently. The latter may be attributed to the fact that numbers are important features to intents, related to doing something on particular dates and with a particular number of people, such as `search_house`, `reserve_restaurant` or `book_appointment`.

7 Conclusion

Over the past years, there has been a trend of utilizing natural language descriptions for various tasks,

ranging from dialog state tracking (Cao and Zhang, 2021), named entity recognition (Li et al., 2020) to the most recent works in text classification employing Pattern-Exploiting Training (PET) (Schick and Schütze, 2020). The help of supervision, expressed in natural language, in most cases not only improves the performance but also enables exploration of real-life setups, such as few-shot or (generalized) zero-shot learning. Such methods’ success is commonly attributed to the efficiency of pre-trained contextualized encoders, which comprise enough prior knowledge to relate the textual task descriptions with the text inputs to the model.

Task-oriented dialogue assistants require the resource-safe ability to support emerging intents without re-training the intent recognition head from scratch. This problem lies well within the generalized zero-shot paradigm. To address it, we present a simple yet efficient approach based on sentence pair modeling, suited for the intent recognition datasets, in which each intent is equipped with a meaningful intent label. We show that we establish new state-of-the-art results using intent labels paired with user utterances as an input to a contextualized encoder and conducting simple binary classification. Besides, to turn intent labels into plausible sentences, better accepted by pre-trained models, we utilized an easy set of lexicalization templates. This heuristic yet alone gains further improvement, increasing the gap to previous best methods. Task transferring from other sentence pair modeling tasks leads to even better performance.

However, our approach has a few limitations: it becomes resource-greedy as it requires to loop over all intents for a given utterance. Next, the intent labels may not be available or may take the form of numerical indices. The first limitation might be overcome by adopting efficient ranking algorithms from the Information Retrieval area. Abstractive summarization, applied to user utterances, might generate meaningful intent labels. These research questions open a few directions for future work.

Acknowledgement

Ekaterina Artemova is supported by the framework of the HSE University Basic Research Program.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ra-

- madan, 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, Brussels, Belgium. Association for Computational Linguistics.
- Jie Cao and Yi Zhang. 2021. [A comparative study on schema-guided dialogue state tracking](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 782–796, Online. Association for Computational Linguistics.
- 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.
- Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.
- 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*.
- Jacek Haneczok, Guillaume Jacquet, Jakub Piskorski, and Nicolas Stefanovitch. 2021. Fine-grained event classification in news-like text snippets shared task 2, case 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).
- 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.
- Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How furiously can colorless green ideas sleep? sentence acceptability in context. *Transactions of the Association for Computational Linguistics*, 8:296–310.
- Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41(5):1202–1241.
- Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified mrc framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859.
- Han Liu, Xiaotong Zhang, Lu Fan, Xuandi Fu, Qimai Li, Xiao-Ming Wu, and Albert Y.S. Lam. 2019a. [Reconstructing capsule networks for zero-shot intent 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 4799–4809, Hong Kong, China. Association for Computational Linguistics.
- 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*.
- Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460.
- Qing Lyu, Hongming Zhang, Elinor Sulem, and Dan Roth. Zero-shot event extraction via transfer learning: Challenges and insights.
- Tingting Ma, Jin-Ge Yao, Chin-Yew Lin, and Tiejun Zhao. 2021. [Issues with entailment-based zero-shot text classification](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 786–796, Online. Association for Computational Linguistics.
- Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353.
- Colin Raffel, Noam 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](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. [Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset](#). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8689–8696. AAAI Press.

- Nils Reimers and Iryna Gurevych. 2019. Sentencebert: 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 3973–3983.
- Oscar Sainz and German Rigau. 2021. Ask2transformers: Zero-shot domain labelling with pre-trained language models. *South African Centre for Digital Language Resources (SADiLaR) Potchefstroom, South Africa*, page 44.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few-shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.
- Qingyi Si, Yuanxin Liu, Peng Fu, Zheng Lin, Jiangnan Li, and Weiping Wang. 2021. Learning class-transductive intent representations for zero-shot intent detection. In *IJCAI*.
- AB Siddique, Fuad Jamour, Luxun Xu, and Vagelis Hristidis. 2021. Generalized zero-shot intent detection via commonsense knowledge. *arXiv preprint arXiv:2102.02925*.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.
- Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 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.
- Chien-Sheng Wu, Steven CH 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.
- Congying Xia, Caiming Xiong, S Yu Philip, and Richard Socher. 2020. Composed variational natural language generation for few-shot intents. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3379–3388.
- Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and S Yu Philip. 2018. Zero-shot user intent detection via capsule neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3090–3099.
- Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 2018. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265.
- Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert YS Lam. 2020. Unknown intent detection using gaussian mixture model with an application to zero-shot intent classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1050–1060.
- Zhiqian Ye, Yuxia Geng, Jiaoyan Chen, Jingmin Chen, Xiaoxiao Xu, SuHua Zheng, Feng Wang, Jun Zhang, and Huajun Chen. 2020. Zero-shot text classification via reinforced self-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3014–3024.
- Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. 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 3905–3914.
- Wenpeng Yin, Nazneen Fatema Rajani, Dragomir Radev, Richard Socher, and Caiming Xiong. 2020. Universal natural language processing with limited annotations: Try few-shot textual entailment as a start. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8229–8239.
- Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

A Reproducibility Checklist

A.1 Code

Our code is enclosed in this submission: `gzsl.zip`.

A.2 Computing infrastructure

Each experiment runs on a single NVIDIA V100 16Gb. The longest experiment was running for less than 2.5 hours.

A.3 Datasets

All used datasets are described in the paper. Preprocessing for SGD and MultiWoZ dataset includes (i) selecting utterances from dialogues where users express a new intent, (ii) cleaning uninformative short phrases like acknowledgments and greetings. Preprocessed datasets are also included in `gzsl.zip`. The SGD dataset is released under CC BY-SA 4.0 license. The MultiWoZ dataset is released under Apache License 2.0. To the best of our knowledge the CLINC dataset is released under CC-BY-3.0 license.

A.4 Randomness

All experiments could be reproduced using the fixed set of seeds $\{11..20\}$.

A.5 Evaluation metrics

All used metrics and our motivation to use them are described in the main paper. Metrics and an evaluation script are implemented in our code.

A.6 Models and hyperparameters

Our sentence pair model consists of the contextualized encoder itself, a dropout, and a linear on top of the embedding for [CLS] token. All hyperparameters for the model are fixed in our submission configs. Transformer tokenizers use truncation for utterance and intent description to speed up execution time. Specified values for lexicalized and non-lexicalized setups are reported in `README.md`.

Batch size, learning rate, scheduler, warm-up steps ratio, and other experiment parameters are specified for each dataset and fixed in configs. We used the top 100 out-of-class similar utterances with a positive one as a threshold for hard negative sampling.

A.7 Hyperparameter Search

We performed hyperparameter search using the following grid for each dataset.

- Learning rate: $[2e^{-5}, 5e^{-5}]$
- Batch size: [8, 16]
- Warm up steps ratio: [0.10, 0.15]
- Utterance max length: [20, 30, 40]
- Negative samples k: [5, 7]

For each hyperparameter configuration, we averaged the results over five runs.

A.8 Acceptability evaluation

Lexicalized intent labels help to increase performance since they form more plausible sentences than raw intent labels. This observation can be confirmed by estimating the acceptability of a sentence. We evaluate the acceptability of intent labels and their lexicalized versions with several unsupervised measures, which aim to evaluate to which degree the sentence is likely to be produced (Lau et al., 2017). We exploit the acceptability evaluation tool from (Lau et al., 2020) with default settings. Following acceptability measures have been used: LP stands for unnormalized log probability of the sentence, estimated by a language model. LP_{mean} and LP_{pen} are differently normalized versions of LP with respect to the sentence length. LP_{norm} and $SLOR$ utilize additional normalization with unigram probabilities, computed over a large text corpus. In this experiment, $BERT_{large}$ is used as the default language model; unigram probabilities are pre-computed from bookcorpus-wikipedia. Higher acceptability scores stand for the higher likelihood of the sentence. Thus, more plausible and more natural-sounding sentences gain higher acceptability scores.

We apply one of the lexicalization patterns to all intent labels, score each resulting sentence, and average the achieved scores. Tables 7-9 present with the results of acceptability evaluation for the each dataset. As expected, the intent labels gain lower acceptability scores, while lexicalized patterns receive higher acceptability scores. We may treat the acceptability of the pattern as a proxy to its performance since the $SLOR$ value of the poor performing q_2 pattern is lower than for other patterns.

ID	LP	LP_{mean}	LP_{pen}	LP_{norm}	$SLOR$
labels	-34.58	-18.40	-30.32	-1.84	-8.44
d ₁	-43.16	-5.55	-23.50	-0.74	1.97
d ₂	-49.92	-5.68	-25.60	-0.77	1.67
q ₁	-46.71	-5.31	-23.93	-0.71	2.12
q ₂	-43.86	-6.48	-25.48	-0.87	0.98

Table 7: Averaged acceptability scores, computed for the CLINC dataset. Rows stand for intent labels without any changes or lexicalized, using one of the patterns. Higher acceptability scores mean that a sentence is more likely to be grammatical and sound natural. Intent labels less acceptable, while their lexicalized versions form plausible sentences.

ID	LP	LP_{mean}	LP_{pen}	LP_{norm}	$SLOR$
labels	-43.18	-18.45	-36.31	-1.96	-9.01
d ₁	-38.93	-5.45	-22.08	-0.72	2.11
d ₂	-43.00	-5.29	-22.92	-0.71	2.09
q ₁	-41.91	-5.14	-22.32	-0.69	2.31
q ₂	-39.36	-6.44	-23.93	-0.86	1.07

Table 8: Acceptability measures, computed for the SGD dataset

ID	LP	LP_{mean}	LP_{pen}	LP_{norm}	$SLOR$
labels	-41.92	-20.96	-37.06	-2.28	-11.77
d ₁	-31.45	-4.49	-18.07	-0.62	2.71
d ₂	-33.90	-4.24	-18.26	-0.60	2.83
q ₁	-33.73	-4.22	-18.17	-0.59	2.93
q ₂	-31.87	-5.31	-19.62	-0.75	1.78

Table 9: Acceptability measures, computed for the MultiWOZ dataset

Method	Unseen		Seen		Overall	
	Acc	F1	Acc	F1	Acc	F1
SP RoBERTa + random IS	0.687 \pm 0.018	0.716 \pm 0.016	0.916 \pm 0.005	0.922 \pm 0.004	0.884 \pm 0.006	0.886 \pm 0.005
SP RoBERTa + random US	0.677 \pm 0.017	0.707 \pm 0.014	0.919 \pm 0.005	0.932 \pm 0.006	0.885 \pm 0.005	0.893 \pm 0.005
SP RoBERTa + hard IS	0.741 \pm 0.010	0.786 \pm 0.017	0.884 \pm 0.010	0.891 \pm 0.012	0.864 \pm 0.009	0.868 \pm 0.010
SP RoBERTa + hard US	0.698 \pm 0.012	0.732 \pm 0.019	0.917 \pm 0.003	0.925 \pm 0.003	0.887 \pm 0.005	0.893 \pm 0.008
SP RoBERTa-NLI	0.748 \pm 0.026	0.801 \pm 0.028	0.923 \pm 0.004	0.929 \pm 0.003	0.898 \pm 0.005	0.905 \pm 0.005
SP BERT-NLI	0.693 \pm 0.017	0.738 \pm 0.015	0.918 \pm 0.002	0.924 \pm 0.001	0.886 \pm 0.003	0.892 \pm 0.002
SP BART-NLI	0.789 \pm 0.024	0.830 \pm 0.030	0.917 \pm 0.000	0.924 \pm 0.000	0.899 \pm 0.003	0.907 \pm 0.005
SP RoBERTa + d ₁ patterns	0.750 \pm 0.019	0.805 \pm 0.021	0.931 \pm 0.006	0.934 \pm 0.004	0.906 \pm 0.004	0.909 \pm 0.002
SP RoBERTa + d ₂ patterns	0.752 \pm 0.003	0.804 \pm 0.006	0.927 \pm 0.007	0.932 \pm 0.004	0.902 \pm 0.005	0.908 \pm 0.003
SP RoBERTa + q ₁ patterns	0.765 \pm 0.019	0.818 \pm 0.021	0.922 \pm 0.010	0.927 \pm 0.010	0.900 \pm 0.007	0.905 \pm 0.007
SP RoBERTa + q ₂ patterns	0.753 \pm 0.026	0.807 \pm 0.026	0.927 \pm 0.005	0.931 \pm 0.002	0.903 \pm 0.004	0.908 \pm 0.004

Table 10: Ablation study, task transferring and lexicalization patterns for SGD dataset. **Top**: comparison of negative sampling strategies of intent sampling (IS) and utterance sampling (US); **middle**: task transferring from the MNLI dataset, using various fine-tuned models; **bottom**: Comparison of different lexicalization patterns, improving performance of SP RoBERTa.

Method	Unseen		Seen		Overall	
	Acc	F1	Acc	F1	Acc	F1
SP RoBERTa + random IS	0.594 \pm 0.180	0.705 \pm 0.157	0.903 \pm 0.055	0.912 \pm 0.047	0.769 \pm 0.082	0.767 \pm 0.084
SP RoBERTa + random US	0.531 \pm 0.218	0.632 \pm 0.217	0.930 \pm 0.036	0.938 \pm 0.027	0.742 \pm 0.096	0.730 \pm 0.106
SP RoBERTa + hard IS	0.561 \pm 0.177	0.680 \pm 0.136	0.937 \pm 0.024	0.943 \pm 0.016	0.771 \pm 0.083	0.761 \pm 0.091
SP RoBERTa + hard US	0.606 \pm 0.244	0.686 \pm 0.234	0.903 \pm 0.033	0.919 \pm 0.030	0.764 \pm 0.099	0.754 \pm 0.108
SP RoBERTa-NLI	0.669 \pm 0.185	0.758 \pm 0.151	0.943 \pm 0.014	0.948 \pm 0.012	0.808 \pm 0.088	0.806 \pm 0.089
SP BERT-NLI	0.624 \pm 0.231	0.715 \pm 0.197	0.941 \pm 0.011	0.948 \pm 0.010	0.785 \pm 0.103	0.782 \pm 0.105
SP BART-NLI	0.673 \pm 0.174	0.753 \pm 0.143	0.946 \pm 0.012	0.950 \pm 0.010	0.820 \pm 0.079	0.814 \pm 0.086
SP RoBERTa + d ₁ patterns	0.624 \pm 0.231	0.722 \pm 0.175	0.941 \pm 0.011	0.948 \pm 0.010	0.785 \pm 0.103	0.782 \pm 0.105
SP RoBERTa + d ₂ patterns	0.610 \pm 0.219	0.713 \pm 0.201	0.944 \pm 0.013	0.948 \pm 0.011	0.786 \pm 0.095	0.781 \pm 0.104
SP RoBERTa + q ₁ patterns	0.621 \pm 0.208	0.727 \pm 0.174	0.946 \pm 0.010	0.949 \pm 0.010	0.789 \pm 0.097	0.786 \pm 0.101
SP RoBERTa + q ₂ patterns	0.599 \pm 0.212	0.702 \pm 0.188	0.943 \pm 0.020	0.948 \pm 0.015	0.778 \pm 0.094	0.775 \pm 0.097

Table 11: Ablation study, task transferring and lexicalization patterns for MultiWoZ dataset.

Method	Unseen		Seen		Overall	
	Acc	F1	Acc	F1	Acc	F1
SP RoBERTa + random IS	0.639 \pm 0.038	0.731 \pm 0.028	0.894 \pm 0.009	0.903 \pm 0.010	0.768 \pm 0.017	0.760 \pm 0.017
SP RoBERTa + random US	0.658 \pm 0.043	0.735 \pm 0.036	0.942 \pm 0.007	0.903 \pm 0.010	0.791 \pm 0.024	0.816 \pm 0.019
SP RoBERTa + hard IS	0.590 \pm 0.039	0.669 \pm 0.036	0.881 \pm 0.008	0.901 \pm 0.010	0.763 \pm 0.020	0.754 \pm 0.018
SP RoBERTa + hard US	0.661 \pm 0.033	0.742 \pm 0.028	0.946 \pm 0.007	0.954 \pm 0.005	0.794 \pm 0.018	0.789 \pm 0.020
SP RoBERTa-NLI	0.700 \pm 0.040	0.771 \pm 0.031	0.950 \pm 0.004	0.955 \pm 0.003	0.817 \pm 0.020	0.836 \pm 0.015
SP BERT-NLI	0.614 \pm 0.035	0.695 \pm 0.026	0.930 \pm 0.007	0.938 \pm 0.007	0.762 \pm 0.020	0.791 \pm 0.018
SP BART-NLI	0.770 \pm 0.039	0.829 \pm 0.034	0.973 \pm 0.003	0.976 \pm 0.002	0.865 \pm 0.022	0.862 \pm 0.024
SP RoBERTa + d ₁ patterns	0.692 \pm 0.031	0.766 \pm 0.028	0.927 \pm 0.009	0.931 \pm 0.008	0.802 \pm 0.018	0.817 \pm 0.015
SP RoBERTa + d ₂ patterns	0.685 \pm 0.035	0.756 \pm 0.031	0.923 \pm 0.014	0.928 \pm 0.012	0.796 \pm 0.024	0.812 \pm 0.021
SP RoBERTa + q ₁ patterns	0.670 \pm 0.034	0.747 \pm 0.029	0.925 \pm 0.010	0.930 \pm 0.009	0.789 \pm 0.019	0.808 \pm 0.015
SP RoBERTa + q ₂ patterns	0.554 \pm 0.054	0.620 \pm 0.055	0.919 \pm 0.008	0.921 \pm 0.009	0.725 \pm 0.029	0.752 \pm 0.022

Table 12: Ablation study, task transferring and lexicalization patterns for CLINC dataset.

Train data	SGD						MultiWoZ						CLINC					
	Unseen			Seen			Unseen			Seen			Unseen			Seen		
	Acc	F1		Acc	F1		Acc	F1		Acc	F1		Acc	F1		Acc	F1	
intent labels + original utterances	0.687 ± 0.018	0.716 ± 0.016		0.916 ± 0.005	0.922 ± 0.004		0.594 ± 0.180	0.705 ± 0.157		0.903 ± 0.055	0.912 ± 0.047		0.639 ± 0.038	0.731 ± 0.028		0.894 ± 0.009	0.903 ± 0.010	
intent labels + synthetic utterances	0.666 ± 0.019	0.688 ± 0.020		0.746 ± 0.014	0.778 ± 0.014		0.615 ± 0.138	0.642 ± 0.090		0.621 ± 0.101	0.713 ± 0.084		0.580 ± 0.045	0.613 ± 0.040		0.608 ± 0.016	0.654 ± 0.009	

Table 13: Dataless classification. Metrics are reported on seen and unseen intents. Fine-tuning SP-Roberta on synthetic utterances (bottom) shows moderate decline, compared to training on real utterances (top).

Method	SGD						MultiWoZ						CLINC					
	Unseen			Seen			Unseen			Seen			Unseen			Seen		
	Acc	F1		Acc	F1		Acc	F1		Acc	F1		Acc	F1		Acc	F1	
SEG	0.372	0.403		0.613	0.636		0.371	0.414		0.652	0.646		-	-		-	-	
RIDE+PU	0.590	0.573		0.832	0.830		0.569	0.521		0.884	0.885		0.798	0.573		0.908	0.912	
ZSDNN + CTIR	0.603 ± 0.002	0.580 ± 0.003		0.809 ± 0.006	0.878 ± 0.014		0.468 ± 0.185	0.437 ± 0.176		0.827 ± 0.022	0.892 ± 0.035		0.561 ± 0.059	0.493 ± 0.054		0.904 ± 0.031	0.871 ± 0.026	
CapsNet + CTIR	0.567 ± 0.017	0.507 ± 0.026		0.897 ± 0.010	0.912 ± 0.009		0.481 ± 0.174	0.404 ± 0.243		0.903 ± 0.017	0.906 ± 0.026		0.530 ± 0.049	0.572 ± 0.033		0.866 ± 0.014	0.883 ± 0.020	
SP RoBERTa (ours)	0.698 ± 0.012	0.732 ± 0.019		0.917 ± 0.003	0.925 ± 0.003		0.606 ± 0.244	0.686 ± 0.234		0.903 ± 0.033	0.919 ± 0.030		0.661 ± 0.033	0.742 ± 0.028		0.946 ± 0.007	0.954 ± 0.005	
SP RoBERTa + patterns (ours)	0.750 ± 0.019	0.805 ± 0.021		0.931 ± 0.006	0.934 ± 0.004		0.624 ± 0.231	0.722 ± 0.175		0.941 ± 0.011	0.948 ± 0.010		0.692 ± 0.031	0.766 ± 0.028		0.927 ± 0.009	0.931 ± 0.008	

Table 14: Comparison of different methods. SP stands for Sentence Pair modeling approach. SP RoBERTa (ours) shows consistent improvements of F1 across all datasets for seen and unseen intents. The usage of lexicalized patterns improves performance.