NLU++: A Multi-Label, Slot-Rich, Generalisable Dataset for Natural Language Understanding in Task-Oriented Dialogue

Iñigo Casanueva,* **Ivan Vulić,*** **Georgios P. Spithourakis,** and **Paweł Budzianowski**
PolyAI Limited
London, United Kingdom
{inigo,ivan,georgios,pawel}@poly.ai

Abstract

We present NLU++, a novel dataset for natural language understanding (NLU) in task-oriented dialogue (ToD) systems, with the aim to provide a much more challenging evaluation environment for dialogue NLU models, up to date with the current application and industry requirements. NLU++ is divided into two domains (BANKING and HOTELS) and brings several crucial improvements over current commonly used NLU datasets. **1)** NLU++ provides fine-grained domain ontologies with a large set of challenging *multi-intent* sentences, introducing and validating the idea of *intent modules* that can be combined into complex intents that convey complex user goals, combined with finer-grained and thus more challenging slot sets. **2)** The ontology is divided into *domain-specific* and *generic* (i.e., domain-universal) intent modules that overlap across domains, promoting cross-domain reusability of annotated examples. **3)** The dataset design has been inspired by the problems observed in industrial ToD systems, and **4)** it has been collected, filtered and carefully annotated by dialogue NLU experts, yielding high-quality annotated data. Finally, we benchmark a series of current state-of-the-art NLU models on NLU++; the results demonstrate the challenging nature of the dataset, especially in low-data regimes, the validity of 'intent modularisation', and call for further research on ToD NLU.

1 Introduction

Research on task-oriented dialogue (ToD) systems (Levin and Pieraccini, 1995; Young et al., 2002) has become a key aspect in industry: e.g., ToD is used to automate telephone customer service tasks ranging from hospitality over healthcare to banking (Raux et al., 2003; Young, 2010; El Asri et al., 2017). Typical ToD systems still rely on a modular design: (i) *the natural language understanding* (NLU) module maps user utterances into a domain-specific set of intent labels and values (Rastogi et al., 2019; Heck et al., 2020; Dai et al., 2021), followed by (ii) the *policy* module, which makes decisions based on the information extracted by the NLU (Gašić et al., 2012; Casanueva et al., 2017; Lubis et al., 2020; Wang et al., 2020a)

Figure 1: Multi-intent examples from the two domains of the NLU++ dataset: BANKING (top) and HOTELS (middle, bottom), illustrating the two core NLU subtasks of intent detection (ID) and slot labeling (SL) in ToD systems. The extracted information is structured into *intents* and *slots*, the latter having associated *values*.

The NLU module is a critical part of any ToD system, as it must extract the *relevant* information from the user's utterances. The information relevance is denoted by the structured *dialogue domain ontology*, which enables the policy module to make decisions about next system actions. The domain ontology covers the information on 1) *intents* and 2) *slots*, see Figure 1. The former is aimed at extracting general conversational ideas (i.e., the user's intents) and corresponds to the standard NLU task of *intent detection (ID)*; the latter extracts specific *slot values* and corresponds to the NLU task of *slot labeling (SL)* (Gupta et al., 2019).[1]

In order to make the policy operational *and* tractable, NLU should extract only the minimal information required by the policy. Therefore, the ontologies differ for each domain of ToD application and are typically built from scratch for each domain.

*Equal contribution.

[1]Slot labeling is also known under other names such as slot filling or value extraction.

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1998 - 2013

Example	Traditional Intent	Intent Modules
I need to change my restaurant reservation	change_restaurant_booking	change, restaurant, booking
When is my booking for the spa?	when_spa_booking	when, spa, booking
TV is not showing any image	tv_not_working	tv, not_working
Why can't I cancel this standing order?	why_cancel_standing_order_not_working	why, cancel, standing_order, not_working

Table 1: Comparison of "traditional" intent annotations vs *intent module*-based multi-label annotations.

Consequently, this makes domain-relevant NLU data extremely expensive to collect and annotate, and prevents its reusability (Budzianowski et al., 2018). Due to this, NLU research in recent years has heavily focused on very data-efficient models that can effectively operate in low-data regimes. Current state-of-the-art (SotA) NLU models leverage large pretrained language models (PLMs) (Devlin et al., 2019; Liu et al., 2019c; Henderson et al., 2020) and fine-tune them with small task-specific datasets (Larson et al., 2019b; Casanueva et al., 2020; Coucke et al., 2018)

At the same time, the progress in creation of NLU datasets has not kept up with the impressive pace of NLU methodology development. However, designing domain ontologies and NLU datasets is also critical for steering further progress in NLU, both from methodology and application perspective. Put simply, current publicly available NLU datasets do not keep up to date with current industry/application requirements for many reasons. **1)** They are usually crowdsourced by untrained annotators (thus typically optimised for quantity rather than quality), yielding examples with low lexical diversity and prone to annotation errors. **2)** They typically assume one intent per example, and thus enable only much simpler single-label ID experiments; such setups are not realistic in more complex industry settings (see Figure 1 again) and lead to unnecessarily large intent sets. **3)** Their ontologies are tied to specific domains, making it difficult to reuse already available annotated data in other domains. **4)** The complexity of the defined tasks and ontologies is limited; the undesired artefact is that current NLU datasets might overestimate the NLU models' abilities, and are not able to separate models any more performance-wise.[2]

In order to address all these gaps, we introduce NLU++, a novel NLU dataset which provides high-quality NLU data annotated by dialogue experts. NLU++ provides *multi-intent*, *slot-rich* and *semantically varied* NLU data, and is inspired by a number of NLU challenges which ToD systems typically face in production environments. Unlike previous ID datasets, examples are annotated with multiple labels, named *intent modules*[3] (see Table 1), with some examples naturally obtaining even up to 6-7 labels. These labels can be seen as sub-intent annotations, where their combinations yield full intents equivalent to "traditional" intents (Table 1). In addition, NLU++ defines a rich set of slots which are combined with the multi-intent sentences. NLU++ is divided into two domains (BANKING and HOTELS) where the two domain ontologies blend a set of *domain-specific* intents and slots with a set of *generic* (i.e., domain-universal) intents and slots. This design makes a crucial step towards generalisation and data reusability in NLU.

Finally, we run a series of experiments on NLU++ with current SotA ID and SL models, demonstrating the challenging nature of NLU++ and ample room for future improvement, especially in low-data setups. Our benchmark comparisons also demonstrate strong performance and shed new light on the (ability of) recently emerging QA-based NLU models (Namazifar et al., 2021; Fuisz et al., 2022), and warrant further research on ToD NLU. The NLU++ dataset is available at: `github.com/PolyAI-LDN/task-specific-datasets`.

2 Background and Motivation

A Brief History of NLU Datasets. As a core module of ToD systems, NLU has been researched since the early 1990s, when the Airline Travel Information System (ATIS) project was started (Hemphill

[2]For instance, for some standard and commonly used NLU datasets such as ATIS (Hemphill et al., 1990; Xu et al., 2020) and SNIPS (Coucke et al., 2018), the results of SotA models are all in the region of 97-98 F_1, with new models getting statistically insignificant gains which might be due to overfitting to the test set or even some remaining annotation errors.

[3]Henceforth, whenever *intents* are mentioned in the context of NLU++, we will be referring to *intent modules*.

		INTENTS			SLOTS		
Domain	**Number of examples**	**Total**	**Generic**	**Avg. per example**	**Total**	**Generic**	**Avg. per example**
BANKING	2,071	48	26	2.25	13	10	0.46
HOTELS	1,009	40	26	1.52	14	10	1.03
ALL	3,080	62	26	2.01	17	10	0.65

Table 2: Key statistics of the NLU++ dataset.

et al., 1990), consisting of spoken queries on flight-related information.[4] Over the next two decades, very few NLU resources were released.[5]

The lack of ToD NLU resources ended in 2013, with the beginning of the 'dialogue state tracking (DST) era' (Williams et al., 2013; Henderson et al., 2014; Kim et al., 2016). Instead of just classifying each turn of the user, DST deals with keeping track of the user's goal over the entire dialogue history, i.e., all the previous user and system turns. Several datasets where released during the DST challenges, all of them comprising simple intent sets (usually tagged as *dialogue acts*).

In order to adapt to the increasing data requirements of deep learning models, increasingly larger dialogue datasets have been released in recent years (Budzianowski et al., 2018; Wei et al., 2018; Rastogi et al., 2019; Peskov et al., 2019). However, the design of ToD datasets comes with some profound differences to datasets for e.g. machine translation or speech recognition, which affect current ToD datasets. **1)** The domain-specific nature of ToD datasets made the data tied to its ontologies, not allowing data reusability across different domains. **2)** The domain-specific ontologies required a lot of expertise for annotation, therefore many annotation mistakes were made (Eric et al., 2019; Zang et al., 2020). **3)** Collecting datasets of that size is unfeasible for development cycles in production, where new domains and models for them need to be very quickly developed and deployed.

Current NLU Trends, inspired by such production requirements, thus deviate from previous DST-oriented NLU research in two main aspects. First, the models went back to focusing on single-turn utterances, which **1)** simplifies the NLU design and **2)** renders the NLU tasks more tractable.[6] The requirement of fast development cycles also instigated more research on NLU (i.e., ID and SL tasks) in low-data scenarios. This way, systems can be developed and maintained faster by reducing the data collection and annotation effort. In addition, the NLU focus shifted from ontologies with only a handful of simple intents and slots (Coucke et al., 2018) to complex ontologies with much larger intent sets (Larson et al., 2019b; Liu et al., 2019b; Casanueva et al., 2020, *inter alia*).

Inspired by these NLU datasets and empowered by transfer learning with PLMs and sentence encoders (Devlin et al., 2019; Liu et al., 2019a; Henderson et al., 2020), there have been great improvements in single-turn NLU systems recently, especially in low-data scenarios (Coope et al., 2020; Mehri and Eric, 2021; Wu et al., 2020b,a; Krone et al., 2020; Henderson and Vulić, 2021; Namazifar et al., 2021; Dopierre et al., 2021; Zhang et al., 2021a,b).

Current Gaps in NLU Datasets. However, existing NLU datasets are still not up to the current industry requirements. **1)** They use crowdworkers for data collection and annotation, often through simple rephrasings; they thus suffer from low lexical diversity and annotation errors (Larson et al., 2019a). **2)** ID datasets always assume a single intent per sentence, which does not support modern production requirements. **3)** The ontologies of these datasets are very domain-specific (i.e., they thus do not allow data reusability) and narrow (i.e., they tend to overestimate abilities of the current SotA NLU models). **4)** Current NLU datasets do not combine a large set of fine-grained intents (again, with multi-intent examples) and a large set of fine-grained slots, which prevents proper and more insightful evaluations of joint NLU models (Chen et al., 2019; Gangadharaiah and Narayanaswamy, 2019).

[4] Remarkably, ATIS is still considered at present as one of the main go-to datasets in NLU reserach. This is also reflected in the fact that the recent most popular dataset for multilingual dialogue NLU was obtained by simply translating English ATIS to 8 more languages (Xu et al., 2020, MultiATIS++).

[5] We note that some Question Classification (Hovy et al., 2001), Paraphrasing (Dolan and Brockett, 2005) and Semantic Text Similarity(Agirre et al., 2012) datasets could be seen as the seed of modern ID datasets, but were not initially built for that purpose.

[6] While DST is theoretically more accurate, it requires amounts of data that grow exponentially with the number of turns; moreover, rule-based trackers have proven to be on par with the learned/statistical ones and require no data (Wang and Lemon, 2013).

Example	Intents	Domain
I want to change my room reservation	change, booking, room	HOTELS
I want to cancel a booking	cancel, booking	HOTELS
Why can't I amend my restaurant booking?	why, change, restaurant, booking, not_working	HOTELS
I am trying to make a transfer but it doesn't let me	make, transfer_payment, not_working	BANKING
I need to increase my overdraft	change, overdraft, higher	BANKING
Please close my savings account	cancel, account, savings	BANKING
The savings one	savings	BANKING
Make it higher	change, higher	GENERAL
Cancel it	cancel	GENERAL
Don't cancel it	deny, cancel	GENERAL

Table 3: NLU++ examples showing the combinatorial expressiveness of intent modules in the multi-intent setting.

We note that there has been some work on multi-label ID on ATIS, MultiWOZ and DSTC4 as multi-intent datasets; however, their multi-label examples remain very limited, simple, and span a small number of intents (Gangadharaiah and Narayanaswamy, 2019). Further, synthetic multi-intent datasets have been created by concatenating single-intent sentences, but such datasets also do not capture the complexity of true and natural multi-intent sentences (Qin et al., 2020).

3 NLU++ Dataset

The NLU++ dataset has been designed with the aim of addressing some of the major shortcomings of the current NLU datasets. In what follows, we describe the main improvements and new evaluation opportunities offered by NLU++.

3.1 Ontology

NLU++ comprises two domains: BANKING and HOTELS. The former represents a banking services task (e.g., making transfers, depositing cheques, reporting lost cards, requesting mortgage information) and the latter is a hotel 'bell desk' reception task (e.g., booking rooms, asking about pools or gyms, requesting room service). Both domains combine a large set of intents with a rich set of slots, with the ontologies inspired by requirements in production. A large number of intents and slots is shared between the two domains, in an attempt to increase data reusability/transferability. Table 2 provides the main statistics of the NLU++ dataset, while the full ontology is presented in Appendix A.

3.2 Multi-Intent Examples

One of the main contributions of this work is the novel design of the intent space, defined in a highly modular manner that natively supports intent re-combinations and multi-intent annotations[7]. For instance, Table 3 shows several multi-intent examples based on the intent sets (termed *intent modules*) from Table 9 in Appendix A.

This design brings several benefits. **1)** The modular nature of the ontology allows for expressing a much more complex set of ideas through different combinations of intent modules (see Table 3), while reducing the overall size of the intent set compared to previous ID datasets[8] (see Table 1 and Table 5). **2)** It allows for the definition of *partial* intents (e.g., *"The savings one"*). This is crucial in multi-turn interactions, where the user often has to answer disambiguation questions (e.g., *"Which account would you like to close?"*). **3)** The modular approach allows the models to generalise to unseen combinations of intent modules. For instance, if (i) examples with the intents *change* and *booking*, and (ii) examples with the intents *cancel* and *account* exist in the training data, (iii) an unseen example with the intents *cancel* and *booking* could be properly predicted, as all the single intents/modules have already been seen by the ID model[9]. **4)** The design also allows us to distinguish between *domain-specific* versus *generic* intent modules. For example, the module *overdraft* is clearly related to BANKING, but the module *change* is much more generic, likely to occur in several different domains.

Finally, the modular design also allows us to

[7] Zhang et al. (2020) proposed a similar way of annotating existing intent detection datasets, showing performance improvements. However, this approach forced categorising the sub-intents in four predefined factors.

[8] Similar to how sub-word tokenization reduced the size of language model vocabularies while covering a larger set of words (Vaswani et al., 2018)

[9] Note that in single-label ID setups, all possible intent module combinations (i.e. "traditional" intents) must be covered (Bi and Kwok, 2013; Hou et al., 2021), which leads to unnecessarily large intent sets and larger data requirements.

study semantic variation of intent modules. Some intents (e.g., especially the domain-specific ones) can only be expressed in a few ways (e.g. *overdraft*, *direct_debit*, *swimming_pool*), while others can have much more varied surface semantic realisations, (e.g. *make, not_working*). Table 9 in Appendix A provides an estimation of the semantic variability of each intent (module).

3.3 Slots

NLU++ further includes a rich set of 17 slots, defined in Table 10 in Appendix A. Table 4 displays several NLU++ examples where complex combinations of intents and slots occur, showcasing how NLU++ might provide a much more challenging environment for the evaluation of joint ID and SL models in future research.

Following the design of previous standard SL datasets (Hemphill et al., 1990; Coucke et al., 2018; Coope et al., 2020), we provide *span annotations for slots*. On top of of this, to also support training and evaluation of SL models which are not span-based, we also provide *value annotations* (or *canonical values* as named by Rastogi et al. (2019)) for times, dates, and numeric values.

Similarly to intent modules, slots can also be divided into the *generic* ones (e.g. *time*, *date*) and the *domain-specific* ones (e.g *company_name*, *rooms*, *kids*), see Table 10. Again, this distinction allows for the cross-domain reusability of annotated data.

3.4 Data Collection and Annotation

Previous NLU datasets have usually relied on *crowdworkers*, aiming to collect a large number of examples, and typically optimising for quantity over quality. However, even with much simpler ontologies, workers are prone to make annotation mistakes, leading to very noisy datasets (Eric et al., 2019). In addition, when workers are asked to rephrase a sentence, they often change its semantic meaning or tend to provide rephrasings with extremely low lexical variability (Kang et al., 2018).

NLU++ reflects true production requirements and focuses on data quality. Instead of relying on crowdworkers, 4 highly skilled annotators with dialogue and NLP expertise, also familiar with production environments, collected, annotated, and corrected the data. The process started by defining the ontology for BANKING and HOTELS. Then, real user examples were fully anonymised and re-annotated following the defined ontology. Finally, new examples were created in order to cover less frequent intents and slots, aiming at creating realistic and semantically varied sentences with new combinations of intents and slots.

3.5 Comparison with Other NLU Datasets

Aiming to reflect the differences between NLU++ and the most popular ToD NLU datasets, Table 5 compares their general statistics. Since the focus of NLU++ is on curated high-quality data, NLU++ covers a fewer number of examples than the other datasets, but it is evident that NLU++ is the only real multi-intent dataset: it averages 2.01 intents per example with a high standard deviation. In addition, NLU++ is the only dataset that combines a large set of intents with a large set of slots.

In order to asses the quality and diversity of the NLU data, we include two additional metrics: 1) Type-Token Ratio (TTR) (Jurafsky and Martin, 2000) which measures lexical diversity) and *semantic diversity*. Both metrics are computed for the set of examples sharing an intent, weighted by the frequency of that intent[10] and finally averaged over intents. The semantic diversity per intent is computed as follows: (i) sentence encodings, obtained by the *ConveRT* sentence encoder (Henderson et al., 2020),[11] are computed for the set of sentences sharing the same intent; (ii) the centroid of these encodings is then computed; (iii) finally, the average cosine distance from each encoding to the centroid is computed. The overall scores clearly indicate that NLU++ offers a much higher lexical and semantic diversity than previous datasets, which should also render it more challenging for current SotA NLU models.[12]

4 Experiments and Results

In hope to establish NLU++ as a more challenging production-oriented testbed for dialogue NLU, especially in low-data scenarios, we evaluate a series of current cutting-edge models for both NLU tasks: intent detection (§4.1) and slot labeling (§4.2). Our aim is to assess and analyse their performance across different setups, and provide solid baseline reference points for future evaluations on NLU++.

Data Setups. Unless noted otherwise, for both tasks we adopt the standard K-fold cross-validation

[10] Note that ATIS has some intents with a single example: for these intents the TTR score would be 1. Weighting by the intent frequency avoids these intents dominating the metric.

[11] See Appendix B for a short description of ConveRT.

[12] SNIPS also shows high semantic diversity, but this is mostly due to the high frequency of named entities.

Example	Intents	Slots (*Values*)
How much less did I spend on Amazon during the current year?	how_much, less, transfer_payment	date_period (*current year*), company_name (*Amazon*)
Show me all the transactions from Sunday to Monday please	request_info transfer_payment	date_from (*Sunday*), date_to (*Monday*)
Hi there, what I want is setting up a 50£ direct debit with Eon for the next 2 months	greet, make, direct_debit	amount_of_money (*50£*), company_name (*Eon*), date_period (*next 2 months*)
Can I make a reservation for 4 adults in 2 rooms, from the 1st of June to the 7th?	make, booking	adults (*4*), rooms (*2*), date_from (*1st of June*), date_to (*7th*)

Table 4: NLU++ examples combining several intents and slots.

Dataset	Number of examples	Number of intents	Number of slots	Avg. intents per example	Avg. slots per example	Type-token ratio (TTR)	Semantic diversity
ATIS	5,871	18	47	1±0.08	3.3±1.61	0.043	0.202
SNIPS	14,484	7	39	1	2.6±1.05	0.154	0.336
OOS	23,700	151	0	1	0	0.148	0.254
BANKING77	13,083	77	0	1	0	0.125	0.209
NLU++	3,080	62	17	2.01±1.25	0.65±0.95	0.268	0.367

Table 5: Comparison of NLU++ with other popular NLU datasets; ATIS (Hemphill et al., 1990), SNIPS (Coucke et al., 2018), OOS (Larson et al., 2019b) and BANKING77 (Casanueva et al., 2020)

as done e.g. by Liu et al. (2019b). Through such *folding* evaluation, (i) we avoid overfitting to any particular test set and (ii) we ensure more stable results with smaller training and test data (i.e., when simulating low-data regimes typically met in production) through averaging over different folds.[13]

The experiments are run with $K = 20$ (**20-Fold**) and $K = 10$ (**10-Fold**), where we train on 1 fold and evalute on the remaining $K - 1$ folds. These setups simulate different degrees of data scarcity: e.g., the average training fold comprises $\approx$ 100 examples for BANKING and $\approx$ 50 for HOTELS for 20-Fold experiments, and twice as much for 10-Fold experiments. Besides these *low-data training setups*, we also run experiments in a **Large**-data setup, where we train the models on merged 9 folds, and evaluate on the single held-out fold.[14] The key questions we aim to answer with these data setups are: Which NLU models are better adapted to low-data scenarios? How much does NLU performance improve with the increase of annotated NLU data? How challenging is NLU++ in low-data versus large-data scenarios?

Domain Setups. Further, experiments are run in the following domain setups: (i) *single-domain* experiments where we only use the BANKING or the HOTELS portion of the entire dataset; (ii) *both-domain* experiments (termed ALL) where we use the entire dataset and combine the two domain ontologies (see Table 2); (iii) *cross-domain* experiments where we train on the examples associated with one domain and test on the examples from the other domain, keeping only *shared* intents and slots for evaluation. The key questions we aim to answer are: Are there major performance differences between the two domains and can they be merged into a single (and more complex) domain? Is it possible to use examples labeled with generic intents from one domain to boost another domain, effectively increasing reusability of data annotations and reducing data scarcity?

F_1 (micro) is the main evaluation measure in all ID and SL experiments.

4.1 Intent Detection: Experimental Setup

We evaluate two groups of SotA intent detection models: (i) *MLP-Based*, and (ii) *QA-Based* ones.

MLP-Based ID Baselines. Casanueva et al. (2020) and Gerz et al. (2021) have recently shown that, for the ID task, full and expensive fine-tuning of large pretrained models such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019a) is not needed to reach strong ID performance. As an alternative, they propose a much more efficient *MLP-based* approach to intent detection which works on par or

[13] Due to folding, variations in results with different random seeds were negligible, even in lowest-data setups.

[14] Effectively, Large-data experiments can be seen as 10-Fold experiments with swapped training and test data.

even outperforms full fine-tuning on the ID task.[15] In a nutshell, the idea is to use fixed/frozen "off-the-shelf" universal sentence encoders such as ConveRT (Henderson et al., 2020) or Sentence-BERT (Reimers and Gurevych, 2019) models to encode input sentences. A standard multi-layer perceptron (MLP) classifier is then learnt on top of the sentence encodings.

Two core differences to the previous work stem from the fact that we now deal with the multi-label ID task: **1)** to this end, we replace the output *softmax* layer with the *sigmoid* layer; and **2)** we define a threshold θ which determines the final classification: only intents with probability scores $\geq \theta$ are taken as positives. This way, the hyper-parameter θ effectively controls the trade-off between precision and recall of the multi-label classifier.

We comparatively evaluate several widely used state-of-the-art (SotA) sentence encoders, but remind the reader that this decoupling of the MLP classification layers from the fixed encoder allows for a much wider empirical comparison of sentence encoders in future work. The evalauted sentence encoders are: **1) ConveRT** (Henderson et al., 2020), which produces 1,024-dimensional sentence encodings; **2) LaBSE** (Feng et al., 2020) (768-dim); **3) RoBL-1B** (1,024-dim) and **4) LM12-1B** (384-dim) (Reimers and Gurevych, 2019; Thakur et al., 2021). For completeness, we provide brief descriptions of each encoder in our evaluation, along with their public URLs, in Appendix B, and refer the reader to the original work for more details about each sentence encoder.

QA-Based ID Baselines. Another group of SotA ID baselines reformulates the ID task into the (extractive) question-answering (QA) problem (Namazifar et al., 2021; Fuisz et al., 2022). This QA-oriented reformatting then allows for additional specialised QA-tuning of large PLMs. In a nutshell, the idea is to (i) fine-tune the original PLM such as BERT/RoBERTa on readily available large general-purpose QA data such as SQuAD (Rajpurkar et al., 2016), and then (ii) further fine-tune this general QA model with in-domain ID data. This strategy has recently shown very strong performance on single-label ATIS data (Namazifar et al., 2021).

The main 'trick' is to reformat the input ID examples into the following format: *"yes. no. [SENTENCE]"* and pose a question such as: *"is the intent to ask about [INTENT]?"* (see Appendix A for the actual questions associated with each intent, also shared with the dataset). Here, *[SENTENCE]* is the placeholder for the actual input sentence, and *[INTENT]* is the placeholder for a short manually defined text (akin to language modeling prompts (Liu et al., 2021), see again Appendix A) which briefly describes the intent. The QA formulation lends itself naturally to the multi-label ID setup as each 'intent-related' question is posed separately. In other words, for each input example and for each of the L intents in the ontology the QA model must extract *yes* or *no* as the answer, where correct intent labels are the ones for which the answer is *yes*.[16] We note that our work is the first to apply and evaluate the QA approach on multi-label ID.

We experiment with two pretrained language models, both fine-tuned on the SQuAD2.0 dataset (Rajpurkar et al., 2018) before additional QA-tuning on NLU++ examples converted to the aforementioned QA format: **RoBB-QA** uses RoBERTa-Base as the underlying LM, while **ALB-QA** relies on the more compact ALBERT (Lan et al., 2020).

ID: Training and Evaluation. All MLP-based baselines rely on the same training protocol and hyper-parameters in all data and domain setups. The MLP classifier consists of 1 hidden layer of size 512, and is trained via binary cross-entropy loss for 500 epochs with the batch size of 32 and the dropout rate is 0.6. We use the standard AdamW optimizer (Loshchilov and Hutter, 2018) with the learning rate of 0.003 and linear decay; weight decay is 0.02. The threshold θ is set to 0.4.[17]

For QA models, we largely follow Namazifar et al. (2021) and fine-tune all models for 5 epochs, using AdamW; the learning rate of $2e$–5 with linear decay; weight decay is 0; batch size is 32.

[15] Our preliminary results on the NLU++ dataset corroborated these findings from prior work; due to a large number of experiments, we thus opt for this more efficient yet also very effective approach to ID.

[16] For instance, for the input sentence *"I need to increase my overdraft"* from the BANKING domain, we would pose all 48 questions associated with each of the $L = 48$ intents in BANKING, where the QA model should extract *yes* as the answer for intents *change*, *overdraft* and *more_higher_after*, and extract *no* for the remaining 45 intents in BANKING.

[17] These hyper-parameters were selected based on preliminary experiments with a single (most efficient) sentence encoder LM12-1B and training only on Fold 0 of the 10-Fold BANKING setup; they were then propagated without change to all other MLP-based experiments with other encoders and in other setups. We repeated the similar hyper-parameter search procedure for QA-based models, using ALB-QA..

Setup→	BANKING			HOTELS			ALL		
	20-Fold	10-Fold	Large	20-Fold	10-Fold	Large	20-Fold	10-Fold	Large
Sentence Encoder↓	**MLP-Based Baselines**								
ConveRT	58.6	70.2	90.3	52.3	63.1	82.8	58.6	70.2	88.9
LaBSE*	54.8	66.6	88.7	48.9	58.9	82.3	55.4	66.1	87.0
RoBL-1B*	56.8	68.4	87.4	55.2	64.2	81.8	57.3	67.7	86.2
LM12-1B*	59.1	69.0	87.8	53.5	62.8	79.5	58.4	68.2	86.0
QA-Pretrained Model↓	**QA-Based Baselines**								
RoBB-QA*	**80.3**	**85.6**	**93.1**	**67.4**	**73.3**	**86.7**	**79.5**	**84**	**91.8**
AlBB-QA*	76.6	82.1	92.0	60.7	67.2	85.1	75.5	80.8	90.6

Table 6: F_1 scores ($\times 100\%$) of benchmarked state-of-the-art intent detection models on NLU++ in three data setups (see §4.1). We also refer to §4 for the brief descriptions of each sentence encoder (for MLP-based baselines) and the two QA-pretrained models. *All models were retrieved from the HuggingFace model repository (Wolf et al., 2020), with exact model URLs available in Appendix §B and Appendix §C. The overall best-performing model per column is in **bold**, while the best-performing MLP-based model per column is underlined.

ConVEx	20-Fold	10-Fold	Large
BANKING	30.1	40.0	68.1
HOTELS	29.7	40.0	64.5
ALL	34.0	45.2	71.4
QA-Based: RoBB-QA	**20-Fold**	**10-Fold**	**Large**
BANKING	50.5	56.7	70.2
HOTELS	48.1	52.4	70.4
ALL	55.5	53.6	72.1

Table 7: F_1 scores ($\times 100\%$) on the NLU++ SL task for ConVEx (Henderson and Vulić, 2021) and a *QA-Based* approach (Namazifar et al., 2021) across different domains and data setups.

	BANKING →HOTELS	HOTELS →BANKING
MLP-Based		
ConveRT	75.4	65.2
LM12-1B	67.3	49.2
QA-Based		
AlB-QA	76.7	72.7
RoBB-QA	**79.3**	**74.2**

Table 8: F_1 scores of *cross-domain* intent detection experiments, evaluating performance on the set of 26 intents shared by the two domains. *Large*-data setup.

4.2 Slot Labeling: Experimental Setup

For slot labeling, we benchmark two current SotA models: (i) *ConvEx* (Henderson and Vulić, 2021), as a SotA span-extraction SL model and (ii) the QA-based SL model (Namazifar et al., 2021) based on RoBB-QA, which operates similarly to QA-based ID baselines discussed in §4.1, and relies on the same fine-tuning regime as our QA-based ID baselines. Again, we refer the reader to the original work for further details, and provide brief descriptions in Appendix D.

4.3 Results and Discussion

Main results with all the evaluated baselines are summarised in Table 6 (for ID) and Table 7 (SL).

ID: MLP versus QA Models. First, the comparisons among only MLP-based models reveal that **1)** all sentence encoders offer ID performance in similar, reasonably narrow score intervals (e.g., the variations in F_1 scores between all sentence encoders are typically below 4-6 F_1 points in all setups), and **2)** that ConveRT is the best-performing sentence encoder on average, which corroborates findings from prior work on other ID datasets (Casanueva et al., 2020; Wu and Xiong, 2020).

One very apparent and important indication in the reported results is the superiority of QA-based ID models over their MLP-based competitors. QA-based models largely outperform MLP-Based baselines in all domain setups, as well as in all data setups. The gains are visible even in Large-data setups, but the benefits of QA-based ID are immense in the lowest-data 20-Fold setups: e.g., 12 F_1 points over the strongest MLP ID model on HOTELS and 20 F_1 points on BANKING.

Moreover, the use of larger underlying LMs might push the scores with QA even further: using SQuAD-tuned Roberta-Large (RoBL-QA) instead of Base (RoBB) yields further gains – e.g., F_1 rises from 85.6 to 87.8 on 10-Fold BANKING, and similar trends are observed in other low-data setups.

Slot Labeling. In the SL task, the QA-based model also demonstrates its superiority, again with huge gains in low-data 20-Fold and 10-Fold setups, confirming that such QA-based or prompt-based methods (Liu et al., 2021; Gao et al., 2021) are especially well suited for low-data setups. The use of manually defined questions/prompts, which are typically easy to write by humans, combined with the expressive power of QA-based task formatting yields immense gains on low-resource dialogue NLU.

Given these very promising ID and SL results on NLU++, our work also calls for further and more intensive future research on QA-based models for dialogue NLU. However, we note that QA-based ID and SL methods do come with efficiency detriments, especially with larger intent and slot sets: the model must copy the input utterance and run a separate answer extraction for each intent/slot from the set, which is by several order of magnitudes more costly at both training and inference than MLP-based models. A promising future research avenue is thus to investigate combined approaches that could combine and trade off the performance benefits of QA-based models and the efficiency advantages of, e.g., MLP-based ID.

Low-Data vs. Large-Data. We also note that scores on both tasks, as reported in Tables 6-7, leave ample room for improvement in NLU methodology in future work, especially on SL (even in Large-data setups), and in low-data setups.

Cross-Domain Experiments. We also verify potential reusability of annotated data across domains with a simple ID experiment, where we train ID models on BANKING and evaluate on HOTELS, and vice versa. The results are summarised in Table 8. Besides (again) indicating that QA-based models outscore MLP-based ID, the results also suggest that for some *generic* intents it is possible to meet high ID performance without any in-domain annotations. For instance, we observe particularly high scores for highly generic and reusable intent modules such as *change*, *how*, *how_much*, *thank*, *when*, and *affirm*, all with per-intent F_1 scores of ≥ 90. We hope that these preliminary results might inspire similar ontology (re)designs in future work.

5 Conclusion

We have presented NLU++, a novel dataset for task-oriented dialogue (ToD) NLU that overcomes the shortcomings of previous NLU evaluation sets. NLU++ presents a multi-intent and slot-rich ontology, defines generic and domain-specific intents and slots to promote data reusability, and it focuses on the creation of high-quality complex examples and annotations collected by dialogue experts. Experimental results show that NLU++ raises the bar with respect to current NLU benchmarks, helping better discriminate and compare the performance of current state-of-the-art NLU models, particularly in low-data setups. We hope that NLU++ will be valuable in guiding future modeling efforts for ToD NLU, both in academia and in industry.

Limitations and Future Work. This work has shown that a better design of the intent set can improve data reusability. However, the current ontology does not cover generic sets of intents exhaustively, and we acknowledge a (sometimes) fine line between truly generic intents versus intents 'anecdotally' shared by two domains (e.g., *refund*). Further, the boundaries of some generic intents can sometimes be unclear and difficult to annotate, even for expert annotators.[18] Future work should try to ground the set of generic intents.

Further, we believe that span-based annotation might be sub-optimal for canonical values such as *times* and *dates*, where small differences in the span would lead to evaluation errors but would not suppose a problem for the value to be parsed. In addition, separating *time* and *date* intervals in different slots increases the difficulty of the annotations and models need to learn a more conflicting set of slots. Further, NLU++ currently provides fine-grained slots such as *date_from*, *date_to* and *date* to enable more complex scenarios, but such a design might slow down annotation process and make it cumbersome. Future work includes rethinking the SL task for these slots.

Finally, while single-turn NLU is more data-efficient and easier to model, some user utterances only make sense in the presence of context from the previous system utterance. While some previous datasets (Coope et al., 2020) deal with this issue with the help of extra annotations indicating if a slot has been requested, in this work we opt for using *non-contextualised* slots such as *number* and *time* and let the policy handle the contextualisation. However, future work should start looking into NLU datasets composed by *system* + *user* turns.

Acknowledgements

We are grateful to our colleagues in PolyAI for many fruitful discussions and suggestions. We also thank the anonymous reviewers for their helpful feedback and comments on the presentation.

Ethical Considerations

PolyAI Limited is ISO27k-certified and fully GDPR-compliant.

[18] For example, boundaries for intents like *greet*, *why* and *change* are clear, while others such as *make* or *not_working* are more prone to ambiguity and different interpretation.

Before data collection: all the data has been collected by workers of PolyAI Limited and all the annotators are also employees of PolyAI Limited.

During data collection: we did not include any personal information (e.g. personal names or addresses) and all the examples that included any had been fully anonymised or removed from the dataset. All the names in the dataset are created by randomly concatenating names and surnames from the list of the top 10K names from the US registry. Upon collection, the dataset has undergone an additional check by the internal Ethics committee of the company. NLU++ is licensed under CC-BY-4.0.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A pilot on semantic textual similarity. In *Proceedings of *SEM*, pages 385–393.

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

Wei Bi and James Tin-Yau Kwok. 2013. Efficient multi-label classification with many labels. In *Proceedings of ICML 2013*, pages 405–413.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of EMNLP 2018*, pages 5016–5026.

Iñigo Casanueva, Paweł Budzianowski, Pei-Hao Su, Nikola Mrkšić, Tsung-Hsien Wen, Stefan Ultes, Lina Rojas-Barahona, Steve Young, and Milica Gašić. 2017. A benchmarking environment for reinforcement learning based task oriented dialogue management. *CoRR*, abs/1711.11023.

Iñigo Casanueva, Tadas Temcinas, 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. *CoRR*, abs/1902.10909.

Sam Coope, Tyler Farghly, Daniela Gerz, Ivan Vulić, and Matthew Henderson. 2020. Span-ConveRT: Few-shot span extraction for dialog with pretrained conversational representations. In *Proceedings of ACL 2020*, pages 107–121.

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. *CoRR*, abs/1805.10190:12–16.

Yinpei Dai, Hangyu Li, Yongbin Li, Jian Sun, Fei Huang, Luo Si, and Xiaodan Zhu. 2021. Preview, attend and review: Schema-aware curriculum learning for multi-domain dialogue state tracking. In *Proceedings of ACL-IJCNLP 2021*, pages 879–885.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Thomas Dopierre, Christophe Gravier, and Wilfried Logerais. 2021. ProtAugment: Intent detection meta-learning through unsupervised diverse paraphrasing. In *Proceedings of ACL-IJCNLP 2021*, pages 2454–2466.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: A corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of SIGDIAL 2017*, pages 207–219.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tür. 2019. MultiWOZ 2.1: Multi-domain dialogue state corrections and state tracking baselines. *CoRR*, abs/1907.01669.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Gabor Fuisz, Ivan Vulić, Samuel Gibbons, Iñigo Casanueva, and Paweł Budzianowski. 2022. Improved and efficient conversational slot labeling through question answering. *CoRR*, abs/2204.02123.

Rashmi Gangadharaiah and Balakrishnan Narayanaswamy. 2019. Joint multiple intent detection and slot labeling for goal-oriented dialog. In *Proceedings of NAACL-HLT 2019*, pages 564–569.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of ACL-IJCNLP 2021*, pages 3816–3830.

Milica Gašić, Matthew Henderson, Blaise Thomson, Pirros Tsiakoulis, and Steve J. Young. 2012. Policy optimisation of POMDP-based dialogue systems

without state space compression. In *Proceedings of SLT 2021*, pages 31–36.

Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michal Lis, Eshan Singhal, Nikola Mrkšić, Tsung-Hsien Wen, and Ivan Vulić. 2021. Multilingual and cross-lingual intent detection from spoken data. In *Proceedings of EMNLP 2021*.

Arshit Gupta, John Hewitt, and Katrin Kirchhoff. 2019. Simple, fast, accurate intent classification and slot labeling for goal-oriented dialogue systems. In *Proceedings of SIGDIAL 2019*, pages 46–55.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gašić. 2020. TripPy: A triple copy strategy for value independent neural dialog state tracking. In *Proceedings of SIGDIAL 2020*, pages 35–44.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS Spoken Language Systems Pilot Corpus. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '90, pages 96–101.

Matthew Henderson, Pawel Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019a. A repository of conversational datasets. In *Proceedings of the 1st Workshop on Natural Language Processing for Conversational AI*, pages 1–10.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of EMNLP 2020*, pages 2161–2174.

Matthew Henderson, Blaise Thomson, and Jason D. Wiliams. 2014. The Second Dialog State Tracking Challenge. In *Proceedings of SIGDIAL 2014*, pages 263–272.

Matthew Henderson and Ivan Vulić. 2021. ConVEx: Data-efficient and few-shot slot labeling. In *Proceedings of NAACL-HLT 2021*, pages 3375–3389.

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019b. Training neural response selection for task-oriented dialogue systems. In *Proceedings of ACL 2019*, pages 5392–5404.

Yutai Hou, Yongkui Lai, Yushan Wu, Wanxiang Che, and Ting Liu. 2021. Few-shot learning for multi-label intent detection. In *Proceedings of AAAI 2021*, pages 13036–13044.

Eduard Hovy, Ulf Hermjakob, Chin-Yew Lin, et al. 2001. The use of external knowledge in factoid QA. In *Proceedings of TREC 2001*, pages 644–652.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st edition. Prentice Hall PTR, USA.

Yiping Kang, Yunqi Zhang, Jonathan K. Kummerfeld, Lingjia Tang, and Jason Mars. 2018. Data collection for dialogue system: A startup perspective. In *Proceedings of NAACL-HLT 2018: Industry Papers*, pages 33–40.

Seokhwan Kim, Luis Fernando D'Haro, Rafael E. Banchs, Jason Williams, and Matthew Henderson. 2016. The fourth dialog state tracking challenge. In *Proceedings of IWSDS 2016*.

Jason Krone, Yi Zhang, and Mona Diab. 2020. Learning to classify intents and slot labels given a handful of examples. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 96–108.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for self-supervised learning of language representations. In *Proceedings of ICLR 2020*.

Stefan Larson, Anish Mahendran, Andrew Lee, Jonathan K. Kummerfeld, Parker Hill, Michael A. Laurenzano, Johann Hauswald, Lingjia Tang, and Jason Mars. 2019a. Outlier detection for improved data quality and diversity in dialog systems. In *Proceedings of NAACL-HLT 2019*, pages 517–527.

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. 2019b. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of EMNLP-IJCNLP 2019*, pages 1311–1316.

Esther Levin and Roberto Pieraccini. 1995. Chronus, the next generation. In *Proceedings of the ARPA Workshop on Spoken Language Technology*.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in Natural Language Processing. *CoRR*, abs/2107.13586.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of ACL 2019*, pages 4487–4496.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019b. Benchmarking natural language understanding services for building conversational agents. In *Proceedings of IWSDS 2019*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019c. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *Proceedings of ICLR 2018*.

Nurul Lubis, Christian Geishauser, Michael Heck, Hsien-chin Lin, Marco Moresi, Carel van Niekerk, and Milica Gasic. 2020. LAVA: Latent action spaces via variational auto-encoding for dialogue policy optimization. In *Proceedings of COLING 2020*, pages 465–479.

Shikib Mehri and Mihail Eric. 2021. Example-driven intent prediction with observers. In *Proceedings of NAACL-HLT 2021*, pages 2979–2992.

Mahdi Namazifar, Alexandros Papangelis, Gokhan Tur, and Dilek Hakkani-Tür. 2021. Language model is all you need: Natural language understanding as question answering. In *Proceedings of ICASSP 2021*, pages 7803–7807.

Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Youssef, and Mona Diab. 2019. Multi-domain goal-oriented dialogues (multidogo): Strategies toward curating and annotating large scale dialogue data. In *Proceedings of EMNLP-IJCNLP 2019*, pages 4526–4536.

Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu. 2020. AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling. In *Findings of EMNLP 2020*, pages 1807–1816.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of ACL 2018*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for Machine Comprehension of Text. In *Proceedings of EMNLP 2016*, page 2383–2392.

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

Antoine Raux, Brian Langner, Alan W. Black, and Maxine Eskénazi. 2003. LET's GO: Improving spoken dialog systems for the elderly and non-natives. In *Proceedings of EUROSPEECH 2003*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP 2019*, pages 3982–3992.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *In Proceedings of NAACL-HLT 2021*, pages 296–310.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas*, pages 193–199.

Kai Wang, Junfeng Tian, Rui Wang, Xiaojun Quan, and Jianxing Yu. 2020a. Multi-domain dialogue acts and response co-generation. In *Proceedings of ACL 2020*, pages 7125–7134.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020b. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained Transformers. In *Proceedings of NeurIPS 2020*.

Zhuoran Wang and Oliver Lemon. 2013. A simple and generic belief tracking mechanism for the Dialog State Tracking Challenge: On the believability of observed information. In *Proceedings of SIGDIAL 2013*, pages 423–432.

Wei Wei, Quoc Le, Andrew Dai, and Jia Li. 2018. Airdialogue: An environment for goal-oriented dialogue research. In *Proceedings of EMNLP 2018*, pages 3844–3854.

Jason D. Williams, Antoine Raux, Deepak Ramachandran, and Alan W. Black. 2013. The Dialogue State Tracking Challenge. In *Proceedings of SIGDIAL 2013*, pages 404–413.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers:

State-of-the-art natural language processing. In *Proceedings of EMNLP 2020: System Demonstrations*, pages 38–45.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020a. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of EMNLP 2020*, pages 917–929.

Chien-Sheng Wu and Caiming Xiong. 2020. Probing task-oriented dialogue representation from language models. In *Proceedings of EMNLP 2020*, pages 5036–5051.

Di Wu, Liang Ding, Fan Lu, and Jian Xie. 2020b. SlotRefine: A fast non-autoregressive model for joint intent detection and slot filling. In *Proceedings of EMNLP 2020*, pages 1932–1937.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of EMNLP 2020*, pages 5052–5063.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In *Proceedings of IJCAI 2019*, pages 5370–5378.

Steve Young. 2010. Cognitive user interfaces. *IEEE Signal Processing Magazine*.

Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. 2002. The HTK book. *Cambridge University Engineering Department*.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117.

Haode Zhang, Yuwei Zhang, Li-Ming Zhan, Jiaxin Chen, Guangyuan Shi, Xiao-Ming Wu, and Albert Y.S. Lam. 2021a. Effectiveness of pre-training for few-shot intent classification. In *Findings of EMNLP 2021*, pages 1114–1120.

Jianguo Zhang, Trung Bui, Seunghyun Yoon, Xiang Chen, Zhiwei Liu, Congying Xia, Quan Hung Tran, Walter Chang, and Philip Yu. 2021b. Few-shot intent detection via contrastive pre-training and fine-tuning. In *Proceedings of EMNLP 2021*, pages 1906–1912.

Jinghan Zhang, Yuxiao Ye, Yue Zhang, Likun Qiu, Bin Fu, Yang Li, Zhenglu Yang, and Jian Sun. 2020. Multi-point semantic representation for intent classification. In *Proceedings of AAAI 2020*, pages 9531–9538.

A Appendix: Ontology

The complete ontology of NLU++ is provided in Table 9 and Table 10.

B Appendix: Sentence Encoders in Intent Detection Experiments

CONVERT (Henderson et al., 2020) is trained with the conversational response selection objective (Henderson et al., 2019b) on large Reddit data (Al-Rfou et al., 2016; Henderson et al., 2019a), spanning more than 700M *(context, response)* sentence pairs. Thanks to its naturally conversational pretraining objective, it has been shown to be especially well-suited for conversational tasks such as intent detection (Casanueva et al., 2020) and slot labelling (Coope et al., 2020). It outputs 1,024-dim sentence encodings.

- github.com/davidalami/ConveRT

LABSE. Language-agnostic BERT Sentence Embedding (LaBSE) (Feng et al., 2020) adapts pretrained multilingual BERT (mBERT) (Devlin et al., 2019) using a dual-encoder framework (Yang et al., 2019) with larger embedding capacity (i.e., a shared multilingual vocabulary of 500k subwords). While LaBSE is the current state-of-the-art multilingual encoder, it also displays very strong monolingual English performance (Feng et al., 2020). It produces 768-dim sentence encodings.

- huggingface.co/sentence-transformers/LaBSE

ROBL-1B and **LM12-1B** (Reimers and Gurevych, 2019; Thakur et al., 2021) are sentence encoders which fine-tune the pretrained Roberta-Large (ROBL) language model (Liu et al., 2019a) and the 12-layer MiniLM (Wang et al., 2020b), respectively, again using a contrastive dual-encoder framework (Reimers and Gurevych, 2019). The models are fine-tuned on a set of more than 1B sentence pairs: this set comprises various data such as Reddit 2015-2018 comments (Henderson et al., 2019a), Natural Questions (Kwiatkowski et al., 2019), PAQ (question, answer) pairs (Lewis et al., 2021), to name only a few.[19] ROBL-1B outputs 1,024-dim encodings, while LM12-1B produces 384-dim encodings.

We opted for those two models in particular as one represents a class of large sentence encoders (ROBL-1B), and the other is lightweight (LM12-1B), while both display very strong performance in a myriad of sentence similarity and semantic search tasks, see www.sbert.net/docs/pretrained_models.html.

- huggingface.co/sentence-transformers/all-roberta-large-v1
- huggingface.co/sentence-transformers/all-MiniLM-L12-v1

C Appendix: QA-Pretrained Models

We rely on the same SQuAD-tuned language models as Namazifar et al. (2021). **ROBB-QA** can be found online at: https://huggingface.co/deepset/roberta-base-squad2; **ALB-QA** is available at: https://huggingface.co/twmkn9/albert-base-v2-squad2

D Appendix: Slot Labeling Baselines

CONVEX (Henderson and Vulić, 2021) demonstrates strong SL performance, especially in few-shot settings. It is pretrained on a pairwise cloze task extracted from the Reddit examples (Henderson et al., 2019a), and the majority of the pretrained model's parameters in CONVEX are kept frozen during fine-tuning, making it an extremely efficient model. We adopt the suggested hyper-parameters from Henderson and Vulić (2021).

QA-Based: Namazifar et al. (2021) train an extractive QA-based model to extract the spans of the slots from the input user utterance as answers to manually defined natural language questions (one per slot). It follows the same idea as QA-based ID models. We also provide such questions for each slot along with NLU++ for model training and inference: see the questions in Table 10.

[19] In a nutshell, the contrastive fine-tuning task which combines all the heterogeneous datasets is as follows: given a 'query' sentence from each sentence pair, and a set of R randomly sampled negatives plus 1 true positive (the sentence from the same pair), the model should predict which sentence from the set of $R + 1$ sentences is actually paired with the query sentence in the dataset. The full list of all datasets along with the exact model specifications is at: huggingface.co/sentence-transformers/all-roberta-large-v1.

INTENT	DESCRIPTION-QUESTION	DOMAIN	LEXICAL DIVERSITY	CATEGORY
affirm	is the intent to affirm something?	general	medium	General dialogue
deny	is the intent to deny something?	general	medium	acts
dont_know	is the intent to say I don't know?	general	high	
acknowledge	is the intent to acknowledge what was said?	general	medium	
greet	is the intent to greet someone?	general	high	
end_call	is the intent to end call or say goodbye?	general	high	
handoff	is the intent to speak to a human or hand off?	general	high	
thank	is the intent to thank someone?	general	medium	
repeat	is the intent asking to repeat the previous sentence?	general	medium	
cancel_close_leave	is the intent asking about canceling or closing something?	general	high	Actions
change	is the intent to change or modify something?	general	high	
make	is the intent to make, open, apply, set up or activate something?	general	high	
request_info	is the intent to ask or request some information?	general	high	Questions
how	is the intent asking how to do something?	general	medium	
why	is the intent to ask why something happened or needs to be done?	general	medium	
when	is the intent to ask about when or what time something happens?	general	medium	
how_much	is the intent asking about some quantity or how much?	general	medium	
how_long	is the intent asking about how long something takes?	general	medium	
not_working	is the intent asking about something wrong, missing or not working?	general	high	General adjectives
lost_stolen	is the intent asking about something being lost or stolen?	general	medium	
more_higher_after	is the intent to indicate something more, higher, after or increasing?	general	medium	
less_lower_before	is the intent to indicate something less, lower, before or decreasing?	general	medium	
new	is the intent asking about something new?	general	medium	
existing	is the intent asking about something that already exists?	general	medium	
limits	is the intent asking about some sort of limit?	general	medium	
savings	is the intent asking about the savings account?	banking	low	Domain specific
current	is the intent asking about the current account?	banking	low	adjectives
business	is the intent to ask something about the business account?	banking	low	
credit	is the intent asking about something related to credit?	banking	low	
debit	is the intent asking about something related to debit?	banking	low	
contactless	is the intent to ask about contactless?	banking	low	
international	is the intent to ask about something related to international issues?	banking	medium	
account	is the intent asking about some account?	banking	low	Domain specific
transfer_payment	is the intent to ask about something related to a transfer, payment or deposit?	banking	low	nouns/entities
appointment	is the intent to ask about something about an appointment?	banking	medium	
arrival	is the intent to ask about the arrival of something?	banking	medium	
balance	is the intent to ask about balance?	banking	medium	
card	is the intent to ask about something related to a card or cards?	banking	low	
cheque	is the intent to ask about cheque?	banking	low	
direct_debit	is the intent to ask about direct debit?	banking	low	
standing_order	is the intent asking about a standing order?	banking	low	
fees_interests	is the intent to ask about fees or interests?	banking	medium	
loan	is the intent to ask about loans?	banking	low	
mortgage	is the intent asking about mortgage?	banking	low	
overdraft	is the intent to ask about ovedraft?	banking	low	
withdrawal	is the intent to ask about withdrawals?	banking	low	
pin	is the intent to ask something about the pin number?	banking	low	
refund	is the intent to ask about some refund?	banking, hotels	low	
check_in	is the intent to ask about check in?	hotels	medium	
check_out	is the intent to ask about check out?	hotels	medium	
restaurant	is the intent to ask something related to restaurant?	hotels	medium	
swimming_pool	is the intent to ask something related to the swimming pool?	hotels	low	
parking	is the intent to ask something related to parking?	hotels	low	
pets	is the intent to ask something related to pets?	hotels	medium	
accesibility	is the intent to ask something related to accessibility?	hotels	medium	
booking	is the intent to talk about some booking?	hotels	medium	
wifi	is the intent to ask something related to wifi or wireless?	hotels	low	
gym	is the intent to ask something related to gym?	hotels	low	
spa	is the intent to ask something related to spa or beauty services?	hotels	high	
room_ammenities	is the intent to ask something related to some room amenities?	hotels	high	
housekeeping	is the intent to talk about housekeeping issues?	hotels	medium	
room_service	is the intent to talk about room service?	hotels	medium	

Table 9: Intents ontology

SLOT	DESCRIPTION-QUESTION	DOMAIN
date	What is the specific date mentioned in this sentence?	general
date_period	What is the time period in days, months or years mentioned in this sentence?	general
date_from	What is the start date of some period mentioned in this sentence?	general
date_to	What is the end date of some period mentioned in this sentence?	general
time	What is the specific time in the day mentioned in this sentence?	general
time_from	What is the start time of some time period mentioned in this sentence?	general
time_to	What is the end time of some time period mentioned in this sentence?	general
time_period	What is the time period in hours or minutes mentioned in this sentence?	general
person_name	What is the name of a person mentioned in this sentence?	general
number	What is the number without context mentioned in this sentence?	general
amount_of_money	What is the specific amount of money mentioned in this sentence?	banking
company_name	What is the name of some sort of company mentioned in this sentence?	banking
shopping_category	What is the category of some expense mentioned in this sentence?	banking
kids	what is the number of kids mentioned in this sentence?	hotels
adults	what is the number of adults mentioned in this sentence?	hotels
people	What is the number of people mentioned in this sentence?	hotels
rooms	What is the number of rooms mentioned in this sentence?	hotels

Table 10: Slots ontology