Recent Neural Methods on Slot Filling and Intent Classification for Task-Oriented Dialogue Systems: A Survey

Samuel Louvan
University of Trento
Fondazione Bruno Kessler
slouvan@fbk.eu

Bernardo Magnini
Fondazione Bruno Kessler
magnini@fbk.eu

Abstract

In recent years, fostered by deep learning technologies and by the high demand for conversational AI, various approaches have been proposed that address the capacity to elicit and understand user's needs in task-oriented dialogue systems. We focus on two core tasks, slot filling (SF) and intent classification (IC), and survey how neural based models have rapidly evolved to address natural language understanding in dialogue systems. We introduce three neural architectures: *independent models*, which model SF and IC separately, *joint models*, which exploit the mutual benefit of the two tasks simultaneously, and *transfer learning models*, that scale the model to new domains. We discuss the current state of the research in SF and IC, and highlight challenges that still require attention.

1 Introduction

The ability to understand user's requests is essential to develop effective task-oriented dialogue systems. For example, in the utterance "*I want to listen to Hey Jude by The Beatles*", a dialogue system should correctly identify that the user's intention is to give a command to play a song, and that *Hey Jude* and *The Beatles* are, respectively, the song's title and the artist name that the user would like to listen. In a dialogue system this information is typically represented through a *semantic-frame* structure (Tur and De Mori, 2011), and extracting such representation involves two tasks: identifying the correct frame (i.e. *intent classification (IC)*) and filling the correct value for the slots of the frame (i.e. *slot filling (SF)*).

In recent years, neural-network based models have achieved the state of the art for a wide range of natural language processing tasks, including SF and IC. Various neural architectures have been experimented on SF and IC, including RNN-based (Mesnil et al., 2013) and attention-based (Liu and Lane, 2016) approaches, till the more recent transformers models (Chen et al., 2019). Input representations have also evolved from static word embeddings (Mikolov et al., 2010; Collobert and Weston, 2008; Pennington et al., 2014) to contextualized word embeddings (Peters et al., 2018; Devlin et al., 2019). Such progress allows to better address dialogue phenomena involving SF and IC, including context modeling, handling out-of-vocabulary words, long-distance dependency between words, and to better exploit the synergy between SF and IC through joint models. In addition to rapid progresses in the research community, the demand for commercial conversational AI is also growing fast, shown by a variety of available solutions, such as Microsoft LUIS, Google Dialogflow, RASA, and Amazon Alexa. These solutions also use various kinds of semantic frame representations as part of their framework.

Motivated by the rapid explosion of scientific progress, and by unprecedented market attention, we think that a guided map of the approaches on SF and IC can be useful for a large spectrum of researchers and practitioners interested in dialogue systems. The primary goal of the survey is to give a broad overview of recent neural models applied to SF and IC, and to compare their performance in the context of task-oriented dialogue systems. We also highlight and discuss open issues that still need to be addressed in the future. The paper is structured as follows: Section 2 describes the SF and IC tasks, commonly used datasets and evaluation metrics. Section 3, 4, and 5 elaborate on the progress and state

Proceedings of the 28th International Conference on Computational Linguistics, pages 480–496
Barcelona, Spain (Online), December 8-13, 2020

of the art of *independent*, *joint*, and *transfer learning models* for both tasks. Section 6 discusses the performance of existing models and open challenges.

Utterance	I	want	to	listen	to	Hey	Jude	by	The	Beatles
Slot	O	O	O	O	O	B-SONG	I-SONG	O	B-ARTIST	I-ARTIST
Intent	PLAY_SONG									

Table 1: Example of SF and IC output for an utterance. Slot labels are in BIO format: B indicates the start of a slot span, I the inside of a span while O denotes that the word does not belong to any slot.

2 Slot Filling and Intent Classification

This section provides some background relevant for SF and IC, sets the scope of the survey with respect to the context in dialogue systems, defines SF and IC as tasks, and introduces the datasets and the metrics that will be used in the rest of the paper.

2.1 Background

Task-oriented dialogue systems aim to assist users to accomplish a task (e.g. booking a flight, making a restaurant reservation and playing a song) through dialogue in natural language, either in a spoken or written form. As in most of the current approaches, we assume a system involving a pipeline of components (Young et al., 2010), where the user utterance is first processed by an Automatic Speech Recognition (ASR) module and then processed by a Natural Language Understanding (NLU) component, which interprets the user's needs. Then a Dialogue State Tracker (DST) accumulates the dialogue information as the conversation progresses and may query a domain knowledge base to obtain relevant data. A dialogue policy manager then decides what is the next action to be executed and a Natural Language Generation (NLG) component produces the actual response to the user.

We focus on the NLU component, and we generalize several recent approaches assuming that the output of the NLU process is a partially filled semantic frame (Wang et al., 2005; Tur and De Mori, 2011), corresponding to the intent of the user in a certain portion of the dialogue, with a number of slot-value pairs that need to be filled to accomplish the intent. The notion of *intent* originates from the idea that utterances can be assigned to a small set of *dialogue acts* (Stolcke et al., 2000), and it is now largely adopted to identify a task or action that the system can execute in a certain domain. *Slot-value pairs*, on the other end, represent the domain of the dialogue, and have been actually implemented either as an ontology (Bellegarda, 2013), possibly with reasoning services (e.g. checking the constraints over slot values) or simply trough a list of entity types that the system needs to identify during the dialogue.

Intents may correspond either to specific needs of the user (e.g. blocking a credit card, transferring money, etc.), or to general needs (e.g. asking for clarification, thanking, etc.). Slots are defined for each intent: for instance, to block a credit card it is relevant to know the name of the owner and the number of the card. Values for the slots are collected through the dialogue, and can be expressed by the user either in a single turn or in several turns. At each user turn in the dialogue the NLU component has to determine the intent of the user utterance (*intent classification*) and has to detect the slot-value pairs referred in the particular turn (*slot filling*). Table 1 shows the expected NLU output for the utterance "*I want to listen to Hey Jude by The Beatles*".

2.2 Scope of the Survey

In Section §2.1, we described a task-oriented system as a pipeline of components, saying that SF and IC are core tasks at the NLU level. Particularly, IC consists of classifying an utterance with a set of predefined intents, while SF is defined as a sequence tagging problem (Raymond and Riccardi, 2007; Mesnil et al., 2013), where each token of the utterance has to be tagged with a slot label. In this scenario training data for SF typically consist of single utterances in a dialogue where tokens are annotated with a predefined set of slot names, and slot values correspond to arbitrary sequences of tokens. In this perspective, it is worth mentioning a research line on dialogue state tracking (Henderson et al., 2014; Mrksic et al., 2015; Budzianowski et al., 2018), where the NLU component is usually embedded into DST. What is

relevant for our topic is that in this context SF is defined as a classification problem: given the current utterance and the previous dialogue history, the system has to decide whether a certain slot-value pair defined in the domain ontology is referred or not in the current utterance. Although promising, from the NLU perspective, this research line poses constraints (e.g. all slot-value pairs have to be pre-defined in an ontology,) that limit the SF applicability. For this reason, and because NLU components are the prevalent solution in current task-oriented systems, the focus of our survey will be on SF as a sequence tagging problem, as more precisely defined in the next section.

2.3 Task Definition

We formulate SF and IC as follows. Given an input utterance $\boldsymbol{x} = (x_1, x_2, .., x_T)$, SF consists in a token-level sequence tagging, where the system has to assign a corresponding slot label $\boldsymbol{y}^{slot} = (y_1^{slot}, y_2^{slot}, .., y_T^{slot})$ to each token x_i of the utterance. On the other end, IC is defined as a classification task over utterances, where the system has to assign the correct intent label y^{intent} for the whole utterance $\boldsymbol{x}$. In general, most machine learning approaches learn a probabilistic model to estimate $p(y^{intent}, \boldsymbol{y}^{slot}|\boldsymbol{x}, \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the parameter of the model. Table 1 shows an example of the expected output of a model for the SF and IC tasks. In the following sections, we outline the main models that have been proposed for SF and IC, and categorize the models into three groups, namely *independent models* (§3) , *joint models* (§4), and *transfer learning based models* (§5).

2.4 Datasets for SF and IC

In this section, according to our task definition, we list available dialogue datasets (most of them are publicly available) where each utterance is assigned to one intent, and tokens are annotated with slot names. Most of such datasets are collections of *single turn user utterances* (i.e., not multi-turn dialogues). An example of a single-turn utterance annotation is shown in Table 1.

The ATIS (Airline Travel Information System) dataset (Hemphill et al., 1990) is the most widely used single-turn dataset for NLU benchmarking. The total number of utterances is around 5K utterances that consist of queries related to the airline travel domain, such as searching for a flight, asking for flight fare, etc. While it has a relatively large number slot and intent labels, the distribution is quite skewed; more than 70% of the intent is a flight search. The slots are dominated by a slot that expresses location names such as FROMLOCATION and TOLOCATION. The MEDIA dataset (Bonneau-Maynard et al., 2005) is constructed by simulating the conversation between a tourist and a hotel representative in the French language. Compared to ATIS, the MEDIA corpus size is around three times larger; however, MEDIA is only annotated with slot labels. The slots are related to hotel booking scenarios such as the number of people, date, hotel facility, relative distance, etc. The MIT corpus (Liu et al., 2013) is constructed through a crowdsourcing platform where crowd workers are hired to create natural language queries in English and annotate the slot label in the queries. The MIT corpus covers two domains, namely movie and restaurant, in which the utterances are related to finding information of a particular movie or actor, searching or booking a restaurant with a particular distance and cuisine criteria. The SNIPS dataset (Coucke et al., 2018) was collected by crowdsourcing through the SNIPS voice platform. Intents include requests to a digital assistant to complete various tasks, such as asking the weather, playing a song, book a restaurant, asking for a movie schedule, etc. SNIPS is now often used as a benchmark for NLU evaluations.

While most datasets are available in English, recently there has been growing interest in expanding slot filling and intent classification datasets to non-English languages. The original ATIS dataset has been derived into several languages, namely Hindi, Turkish (Upadhyay et al., 2018), and Indonesian (Susanto and Lu, 2017). The MultiATIS++ dataset from Xu et al. (2020) expands the ATIS dataset to more languages, namely Spanish, Portuguese, German, French, Chinese, and Japanese. The work from (Bellomaria et al., 2019) introduces the Italian version of the original SNIPS dataset. The Facebook multi-lingual dataset (Schuster et al., 2019), introduced a dataset on Thai and Spanish languages across three domains namely weather, alarm, and reminder. The detailed statistics of each dataset are listed in Appendix A.

2.5 Evaluation Metrics

For the IC task, evaluation is performed on the utterance level. The typical evaluation metric for IC is *accuracy*, calculated as the number of the correct predictions made by the model divided by the total number of predictions. As for SF, the evaluation is performed on the entity level. The common metrics used is the metric introduced in CoNLL-2003 shared task (Sang and Meulder, 2003) to evaluate Named Entity Recognition (NER) by computing the F-1 score. The F1-score, is the harmonic mean score between precision and recall. Precision is the percentage of slot predictions from the model which are correct, while recall is the percentage of slots in the corpus that are found by the model. A slot prediction is considered *correct* when an *exact* match is found (Sang and Meulder, 2003). As the slot is annotated in BIO format to mark the boundary of the slot (see Table 1), a correct prediction is only counted when the model can predict the correct slot label on the correct token offset. Consequently, the exact match metrics does not reward cases when the model predict correct slot label but get the incorrect slot boundary (*partial match*).

3 Independent Models for SF and IC

Independent models train each task *separately* and recent neural models typically use RNN as the building block for SF and IC. At each time step t, the encoder transforms the word representation x_t to the hidden state h_t. For SF, the output layer predicts the slot label y_t^{slot} condition on h_t. For IC, typically the last hidden state h_T is used to predict the intent label $y^{intent}$of the utterance $\boldsymbol{x}$. Note that, for independent approaches, the models for SF and IC are trained separately. Most neural models for SF and IC generally consist of several layers, namely an *input layer*, one or more *encoder layer*, and an *output layer*. Consequently, the main differences between models are in the specifics of these layers. The most common dataset used for evaluating independent models is ATIS.

In the *input layer* of neural models each word is mapped into embeddings. Mesnil et al. (2013) compared several embeddings, namely pre-trained SENNA (Collobert et al., 2011), RNN Language Model (RNNLM) (Mikolov et al., 2011), and random embeddings. SENNA gives the best result compared to other embeddings, and, typically, further fine-tuning word embeddings improves performance. (Yao et al., 2013) report that embeddings learned from scratch directly on ATIS data (*task-specific embeddings*) are better than SENNA. However, task-specific embeddings are composed not only by words but also by named entities (*NE*) and syntactic features[1]. NE improves performance significantly while part-of-speech only adds small benefits. Ravuri and Stolcke (2015) emphasizes the importance of *character representation* to handle OOV issues.

For the *encoder layer*, various RNN architectures have been applied to SF and IC (Mesnil et al., 2013; Mesnil et al., 2015; Liu and Lane, 2015). Mesnil et al. (2013) compare the Elman (Elman, 1990) and Jordan (Jordan, 1997) RNNs. They observe that the performance of the Jordan RNN is marginally better than Elman. They also experiment a *bi-directional* version of Jordan RNN and obtained the best score of 93.89 F1 for SF, performing better than CRF for about +1 absolute F1 improvement. Xu and Sarikaya (2013) use Convolutional Neural Network (CNN) (LeCun et al., 1998) to extract 5-gram features and apply max-pooling to obtain the word representation before passing it to the output layer. Compared with RNN (Yao et al., 2013; Mesnil et al., 2013), CNN gives lower performance for SF on ATIS. Other studies (Yao et al., 2014a; Vu et al., 2016) adapt Long Short-Term Memory Network (LSTM) (Hochreiter and Schmidhuber, 1997) to SF. The LSTM model gives better SF performance compared to CRF, CNN, and RNN. Ravuri and Stolcke (2015) compare the performance of vanilla RNN and LSTM for IC. They find that the vanilla RNN works best for shorter utterances, while LSTM is better for longer utterances.

For the *output layer*, typically a *softmax* function is used for prediction at a particular time step. Yao et al. (2014b) propose a R-CRF model combining the feature learning power of RNN and the *sequence level optimization* of CRF for SF. The RNN + CRF scoring mechanism incorporates the features learned from RNN and the transition scores of the slot slot labels. R-CRF outperforms CRF and vanilla RNN on ATIS and on the Bing query understanding dataset. Table 2 summarizes the performance of independent models on SF and IC.

[1]Gold named entity and syntactic information

	Input	Model (Enc/Dec)	Output	Slot (F1)	Intent(Err)
Xu and Sarikaya (2013)	lexical	CNN	softmax	94.35	6.65
Yao et al. (2013)	lexical	Elman RNN	softmax	94.11	-
Yao et al. (2013)	lexical+NE	Elman RNN	softmax	96.60 -	
Yao et al. (2014a)	lexical	LSTM	softmax	94.85	-
Yao et al. (2014b)	lexical+NE	Elman RNN	CRF	96.65	-
Mesnil et al. (2015)	lexical	Hybrid Elman + Jordan RNN	softmax	95.06	-
Liu and Lane (2015)	lexical	Elman RNN with label sampling	softmax	94.89	-
Vu et al. (2016)	lexical	bi-directional RNN	softmax	94.92	-
Liu and Lane (2016)	lexical	bi-directional RNN+attention	softmax	95.75	2.35
Kurata et al. (2016)	lexical	Encoder-Decoder LSTM	softmax	95.40	-

Table 2: Comparison of independent SF and IC models and their performance on ATIS.

Takeaways on independent SF and IC models:

- Performance of RNN encoders (*unidirectional*) are Jordan $\leq$ Elman $<$ LSTM. Bi-directional encoding is additive to the performance of each encoder.
- Incorporating more context information is better for SF performance. Using global context information, such as sentence level representation, and attention mechanisms (Kurata et al., 2016; Liu and Lane, 2016) boosts performance of bi-directional encoder even further.
- When adding external features is possible, semantic features such as NE are more beneficial than syntactic features for SF. When NE is used, it can boost the model performance for SF significantly.
- The slot filling task is related to Named Entity Recognition (NER) (Grishman and Sundheim, 1996) task as slot values can be a named entity such as airline name, city name etc. If the slot filling task is modeled as a sequence tagging problem, basically recent neural models proposed for NER can be used for slot filling and vice versa. To know more about the recent development of neural NER models, one can consult the survey from Yadav and Bethard (2018).
- The main disadvantage of independent models is that they do not exploit the interaction between intent and slots and may introduce error propagation when they are used in a pipeline.

4 Joint Models for SF and IC

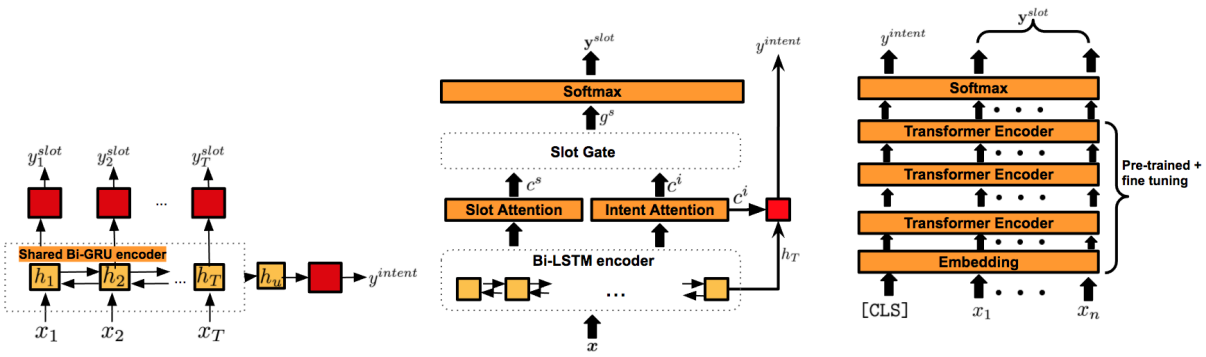

Figure 1: *Left:* Shared Bi-GRU encoder (Zhang and Wang, 2016). *Middle:* Slot-Gate Mechanism (Goo et al., 2018). *Right:* BERT Based (Chen et al., 2019).

In Section 3 we reported approaches that treat SF and IC *independently*. However, as the two tasks always appear together in an utterance and they share information, it is intuitive to think that they can benefit each other. For instance, if the word "*The Beatles*" is recognized as the slot ARTIST, then it is more likely that the intent of the utterance is PLAYSONG rather than BOOKFLIGHT. On the other hand, recognizing that the intent is PLAYSONG would help to recognize "*Hey Jude*" as the slot ARTIST rather than MOVIENAME.

Recent approaches model the relationship between SF and IC *simultaneously* in a *joint model*. These approaches promote *two-way* information sharing between the two tasks instead of a one-way (*pipeline*).

We describe several alternatives to exploit the relation between SF and IC: through *parameter and state sharing* and *gate mechanism*.

4.1 Parameter and State Sharing

A pioneering work in joint modeling is Xu and Sarikaya (2013), which performs parameter sharing and captures the relation between SF and IC through Tri-CRF (Jeong and Lee, 2008). The model uses CNN as a *shared* encoder for both tasks and the produced hidden states are utilized for SF and IC. In addition to features learned from the NN and from the slot label transition, Tri-CRF incorporates an additional factor g to learn the correlation between the slot label assigned to each word and the intent assigned to the utterance, which explicitly captures the dependency between the two tasks. A similar approach (Guo et al., 2014), shares the node representation produced by Recursive Neural Network (RecNN) which operates on the syntactic tree of the utterance. The node's representation is *shared* among SF and IC. Zhang and Wang (2016) use a *shared* bi-GRU encoder and a *joint loss function* between SF and IC (Figure 1 *Left*), in which the loss function has weights associated with each tasks.

Liu and Lane (2016) use a neural sequence to sequence (encoder-decoder) model with attention mechanism commonly used for neural machine translation. The *shared* encoder is a bi-directional LSTM, and the last hidden state of the encoder is then used by the decoder to generate a sequence of slot labels, while for IC there is a separate decoder. The attention mechanism is used to learn alignments between slot labels in the decoder and words in the encoder. Hakkani-Tür et al. (2016) also adopt parameter sharing similar to Zhang and Wang (2016), but instead of using GRU they use a shared LSTM and perform predictions for slots, intent, and also domain.

In a recent approach by Wang et al. (2018) propose a bi-model based structure to learn the *cross-impact* between SF and IC. They argue that a single model for two tasks can hurt performance, and, instead of sharing parameters, they use two-task networks to learn the cross-impact between the two tasks and only share the hidden state of the other task. In the model, every hidden state h_t^1 in the first network is combined with the hidden state of the second network h_t^2, and vice versa. Training is also done asynchronously, as each model has a separate loss function. Qin et al. Qin et al. (2019) use a self-attentive shared encoder to produce better context-aware representations, then apply IC at the *token level* and use this information to guide the SF task. They argue that previous work based on *single utterance-level* intent prediction is more prone to error propagation. If some token-level intent is incorrectly predicted, the other correct token-level prediction can still be useful for corresponding SF. For the final IC prediction, they use a voting mechanism to take into account the IC prediction on each token.

Chen et al. (2019) use a Transformer (Vaswani et al., 2017) model for joint SF and IC by fine-tuning a pre-trained BERT (Devlin et al., 2019) model (Figure 1 *Right*). The input is passed through several layers of transformer encoders and the hidden state outputs are used to compute slot and intent labels. The hidden state $h^{\texttt{CLS}}$ is used for IC[2] while the rest of the hidden states at each time step h_i serve SF.

4.2 Slot-Intent Gate Mechanism

In addition to parameter and state sharing, a separate network with a *slot gating mechanism* was introduced by Goo et al. (2018) to model the interaction between SF and IC more explicitly (Figure 1 *Middle*). In the encoder, a *slot context vector* for each time step, $\boldsymbol{c}_i^S$, and a global intent context vector $\boldsymbol{c}^I$ are computed using an attention mechanism (Bahdanau et al., 2015). The slot-gate $\boldsymbol{g}^s$ is computed as a function of $\boldsymbol{c}_i^S$ and $\boldsymbol{c}^I$, $\boldsymbol{g}^s = \sum v \cdot tanh(\boldsymbol{c}_i^S + \boldsymbol{W} \cdot \boldsymbol{c}^I)$. Then, $\boldsymbol{g}^s$ is used as a weight between $\boldsymbol{h}_i$ and $\boldsymbol{c}_i^S$ to compute y_i^{slot} as follows: $y_i^{slot} = \texttt{softmax}(\boldsymbol{W}(\boldsymbol{h}_i + \boldsymbol{g}^s \cdot \boldsymbol{c}_i^S))$. Larger $\boldsymbol{g}^s$ indicates a stronger correlation between c_i^S and c^I.

E et al. (2019) propose a bi-directional model, SF-ID (SF-Intent Detection) network, sharing ideas with Goo et al. (2018), with two key differences. First, in addition to the slot-gated mechanism, they add an intent-gated mechanism as well. Second, they use an iterative mechanism between the SF and ID network, meaning that the gate vector from SF is injected into the ID network and vice versa. This

[2] `[CLS]` is a special token in BERT input format that often used as the sentence representation.

mechanism is repeated for an arbitrary number of iteration. Compared to (Goo et al., 2018), the SF-ID network performs better both in SF and IC on ATIS and SNIPS. The work from Li et al. (2018) is also similar to Goo et al. (2018) with two differences. First, they use a self-attention mechanism (Vaswani et al., 2017) to compute c_i^S. Secondly, they use a separate network to compute gate vector g^s, but the input of this network is the concatenation of c_i^S and the intent embedding v, and g^s is defined as $g^s = \texttt{tanh}(\boldsymbol{W}^g[c_{slot}^i, v^{intent}] + b^s)$. After that, h_i is combined with g^s through element-wise multiplication to compute y_i^s as follows: $y_i^{slot} = \texttt{softmax}(\boldsymbol{W}^s(h_i \odot g^s) + b^s)$. They report a +0.5% improvement on SF over Liu and Lane (2016). A recent work by Zhang et al. (2019), further improves the performance of the BERT based model by adding a gate mechanism (Li et al., 2018) to the BERT model. Table 3 compares the performance of the joint models.

Method	Model	ATIS		SNIPS	
		Slot F1	Intent Acc/Err	Slot F1	Intent Acc/Err
Parameter & State Sharing					
Xu and Sarikaya (2013)	CNN + Tri-CRF	95.42	-/5.91	-	-
Guo et al. (2014)	Recursive NN	93.96	95.40	-	-
Zhang and Wang (2016)	Joint Multi-Task,Bi-GRU	95.49	98.10	-	-
Liu and Lane (2016)	Seq2Seq + Attention	94.20	91.10	87.80	96.70
Hakkani-Tür et al. (2016)	Bi-LSTM	94.30	92.60	87.30	96.90
Qin et al. (2019)	Token-Level IC + Self-Attention	95.90	96.90	94.20	98.00
Chen et al. (2019)	Transformer (BERT)	96.10	97.50	97.00	98.60
State Sharing					
Wang et al. (2018)	Bi-model, BiLSTM	96.89	98.99	-	-
Slot-Intent Gating					
Goo et al. (2018)	Slot-Gated Full Attention	94.80	93.60	88.80	97.70
Li et al. (2018)	BiLSTM + Self-Attention	96.52	-/1.23	-	-
E et al. (2019)	SF-ID Network	95.75	97.76	91.43	97.43
Hybrid Param Sharing + Gating					
Zhang et al. (2019)	BERT + Intent-Gate	98.75	99.76	98.78	98.96

Table 3: Performance comparison of joint models for SF and IC on ATIS and SNIPS-NLU.

Takeaways on joint SF and IC models:

- The overall performance of joint models for SF and IC (Table 2) is competitive with independent models (Table 3). The advantage of joint models is that they have relatively less parameters than independent models, as both tasks are trained on a single model.
- When computational power is not an issue, fine-tuning a pre-trained model such as BERT is the way to go for maximum SF and IC performance. Hybrid methods combining parameter and state sharing + intent gating yield the best performance (Zhang et al., 2019).
- For the non BERT-based model, using state sharing (Wang et al., 2018) is the best on ATIS. However, the disadvantage is that it is actually a bi-model and not a single model.
- Similar to independent models, contextual information is crucial to performance. Adding a self-attention mechanism (Qin et al., 2019; Li et al., 2018) to either parameter and state sharing or to slot-intent gating can boost performance even further.
- When sufficiently large in-domain training data is available, the SF and IC performance in ATIS and SNIPS is already saturated. Therefore, further research on this classic leaderboard chase is not worth it. We discuss more about that in Section 6.
- Most of the work in joint models and also independent models (Section §3) reports F1 scores for slot filling performance. However, these scores do not reveal in which specific cases these models behave differently, contributing to overall performance. We leave further analysis on model performance as a potential future work.

5 Scaling to New Domains

So far, the models that we consider in Section §3 and Section §4 are designed to be trained on a *single domain* (e.g. banking, restaurant reservation) and require relatively *large labeled data* to perform well. In practice, new intents and slots are regularly added to a system to support new tasks and domains, requiring data and time intensive processes. Hence, methods to train models for new domains with limited or without labeled data are needed. We refer to this situation as the *domain scaling* problem.

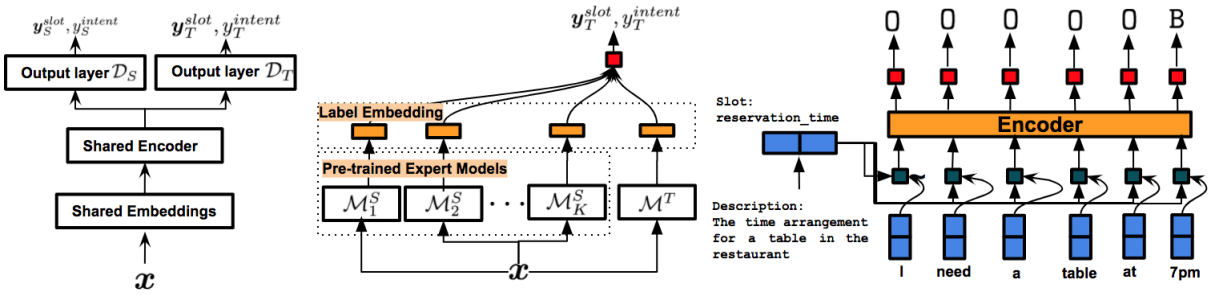

Figure 2: *Left:* Data-driven approach (Jaech et al., 2016; Hakkani-Tür et al., 2016). *Middle:* Model-Driven Approach with expert models (Kim et al., 2017). *Right:* Zero-shot model (Bapna et al., 2017).

5.1 Transfer Learning Models for SF and IC

A common approach to deal with domain scaling is transfer learning (TF).[3] In the TF setup we have K source domains $\mathcal{D}_S^1, \mathcal{D}_S^2, \ldots, \mathcal{D}_S^K$ and a target domain $\mathcal{D}_T^{K+1}$, and we assume abundance of data in $\mathcal{D}_S$ and limited data in $\mathcal{D}_T$. Instead of training a target model $\mathcal{M}_T$ for $\mathcal{D}_T$ from scratch, TF aims to *adapt* the learned model $\mathcal{M}_S$ from $\mathcal{D}_S$ to produce a model $\mathcal{M}_T$ trained on $\mathcal{D}_T$. TF is typically applied with various parameter sharing and training mechanisms. For SF and IC two approaches are proposed, namely *data-driven* and *model- driven*. As for data-driven techniques, typically we combine data from $\mathcal{D}_S$ and $\mathcal{D}_T$ and we partition the parameters in the model into parts that are *task-specific* and parameters that are shared across tasks. Some studies (Jaech et al., 2016; Hakkani-Tür et al., 2016; Louvan and Magnini, 2019) apply this technique using *multi-task learning* (MTL) (Caruana, 1997) and the models are trained simultaneously on $\mathcal{D}_S$ and $\mathcal{D}_T$ (Figure 2 *Left*). Results have shown that MTL is particularly effective relative to single-task learning (STL) when the data in $\mathcal{D}_T$ is scarce and the benefits over STL diminish as more data is available. Another technique that is typically used in data-driven approaches is based on *pre-train* and *fine-tune* mechanisms. Goyal et al. (2018) train a joint model of SF and IC, $\mathcal{M}_S$, on large $\mathcal{D}_S$, then fine-tune $\mathcal{M}_S$ by replacing the output layer corresponding with the label space from $\mathcal{D}_T$ and train the model further on $\mathcal{D}_T$. Siddhant et al. (2019) also uses fine-tuning mechanism, but the main difference with Goyal et al. (2018) is they leverage large unlabeled data to learn contextual embedding, ELMo (Peters et al., 2018), before fine-tuning on $\mathcal{D}_T$.

As we need to train from scratch the whole model when adding a new domain, data-driven approaches, especially MTL-based, need increasing training time as the number of domains grows. The alternative strategy, the model-driven approach, alleviates the problem by enabling model *reusability*. Although different domains have different slot schemas, slots such as *date*, *time* and *location* can be shared. In model driven adaptation "expert" models (Figure 2 *Middle*) are first trained on these reusable slots (Kim et al., 2017; Jha et al., 2018) and the outputs of the expert models are used to guide the training of $\mathcal{M}_T$ for a new target domain. This way the training time of $\mathcal{M}_T$ is faster, as it is proportional to the $\mathcal{D}_T$ data size, instead of the larger data size of the whole $\mathcal{D}_S$ and $\mathcal{D}_T$. In this model-driven settings, Kim et al. (2017) do not treat each expert model on each $\mathcal{D}_S$ equally, instead they use attention mechanism to learn a weighted combinations from the feedback of the expert models. Jha et al. (2018) use a similar model as Kim et al. (2017), however they do not use attention mechanism. For training the expert models, instead of using all available $\mathcal{D}_S$, they build a repository consisting of common slots, such as *date*, *time*, *location*

[3] We do not differentiate between *domain adaptation* and transfer learning in this paper.

slots. The assumption is that these slots are potentially reusable in many target domains. Upon training $\mathcal{M}_S$ on this reusable repository, the output of $\mathcal{M}_S$ is directly used to guide the training of $\mathcal{M}_T$.

5.2 Zero-shot Models for SF and IC

While data-driven and model-driven approaches can share knowledge learned on different domains, such models are still trained on a pre-defined set of labels, and can not handle *unseen* labels, i.e. not mapped to the existing schema. For example, a model trained to recognize a DESTINATION slot, can not be used directly to recognize the slot ARRIVAL_LOCATION for a new domain, although both slots are semantically similar. For this reason, researchers have recently been working on *zero-shot* models, trained on *label representations* that leverage natural language *descriptions* of the slots (Bapna et al., 2017; Lee and Jha, 2019). Assuming that accurate slot descriptions are provided, slots with *different* names although semantically similar would have similar description as well. Thus, having trained a model for the DESTINATION slot with its descriptions, it is now possible to recognize the slot ARRIVAL_LOCATION without training on it, but only supplying the corresponding slot description.

In addition to slot description, other zero-shot approaches explore the use of slot value examples (Shah et al., 2019; Guerini et al., 2018). Shah et al. (2019) showing that a combination of a small number of slot values examples with a slot description performs better than (Bapna et al., 2017; Lee and Jha, 2019) on the SNIPS dataset. Zero-shot models are typically trained on a per-slot basis (Figure 2 *Right*), meaning that if we have N slots, then the model will output N predictions, therefore, a merging mechanism is needed in case there are prediction overlaps. In order to alleviate the problem of having multiple predictions, Liu et al. (2020b) propose a *coarse-to-fine* approach, in which the model learns the slot entity pattern (coarsely) to identify a particular token is an entity or not. After that, the model performs a single prediction of the slot type (fine) based on the similarity between the feature representation and the slot description.

Takeaways on scaling to new domains:

- Both data driven methods, MTL and pre-train fine tuning, improve performance when data in $\mathcal{D}_T$ is limited. Both are also flexible, as virtually many tasks from different domains can be plugged into these methods. As the number of domains grow, pre-train and fine tuning is more desirable than MTL. However, fine tuning is more prone to the *forgetting* problem (He et al., 2019) compared to MTL.
- When the number of domain, K, is massive, the pre-train fine tuning approach and model driven approaches, such as expert based adaptation, are preferable with respect of training time.
- When there exists K existing domains and no annotation is available in $\mathcal{D}_T$, the choice is zero-shot approaches with the expense of providing meta-information such as slot and intent descriptions.
- As typically zero-shot models perform prediction on a *per-slot* basis, potential disadvantages are model accuracy when there is a prediction overlap and the model can also be computationally inefficient when dealing with many slots.

6 State of the Art and Beyond

Based on the results in Table 2 and 3, it is evident that neural models have achieved outstanding performance on ATIS and SNIPS, showing that it is relatively easy for neural models to capture patterns that recognize slots and intents. ATIS, in particular, is already overused for SF and IC evaluations and recent analysis (Béchet and Raymond, 2018; Niu and Penn, 2019) have shown that the dataset is relatively simple and the room for performance improvement is tiny. A similar trend in performance can be noted for other datasets, such as SNIPS, and it is likely that performance improvement can be quickly saturated. However, it does not mean these models have solved SF and IC, or NLU problems in general, rather that the model has merely solved the datasets. Nevertheless, there are still a number of issues in SF and IC that need further investigation:

Portable and Data Efficient Models. Instead of evaluating models with the typical *leaderboard* setup with fixed (train/dev/test) splits on a specific target domain, it would be also important to test models in different scenarios, so that different aspects of the model can be captured. For example, as neural models

are data hungry, more work is still needed on transfer learning scenarios, where evaluation is carried out with *less* or *without* labeled data (*zero-shot*) for a particular target domain. In addition, most models for SF and IC are evaluated on English, which means that more effort is still needed to make models that work well for other languages. Some recent works have started exploring zero-shot cross lingual methods (Qin et al., 2020; Liu et al., 2020a; Liu et al., 2019) and also few-shot scenarios (Hou et al., 2020) and the room for improvement for these scenarios is still large. In short, designing a *data efficient* model that is *portable* across domains and languages is still a challenging problem for the coming future.

Leveraging unlabeled data from live traffic. In real situations, personal digital assistants such as Google Home, Apple Siri and Amazon Alexa, receive live traffic data from real users. This large amount of unlabeled data from live traffic is a potential data source for model training, in addition to in-house annotated data. Unlabeled live data are likely different from in-house data, as they can contain more diverse utterances and also noisy and irrelevant utterances. In this situation, existing methods to tap on unlabeled data, such as semi-supervised learning, still face unique challenges to handle live data. It is worth to note that a bottleneck in this direction is that working on live data in academic settings is not trivial. Some recent works explore this line of research by applying semi-supervised learning (Cho et al., 2019) and also data selection (Do and Gaspers, 2019) mechanism.

Generative Models. Most of the proposed models are *discriminative*, among the few works carried out for *generative models* for SF and IC, (Raymond and Riccardi, 2007; Yogatama et al., 2017) have shown that a generative model is relatively better than a discriminative model in a situation where data is *scarce*. One possible direction for generative models is to apply data augmentation to automatically create additional training data (Yoo et al., 2019; Zhao et al., 2019; Hou et al., 2018; Kurata et al., 2016; Peng et al., 2020; Kim et al., 2019). The main challenge for data augmentation is to generate diverse and fluent synthetic utterance, which *preserve* the semantics of the original utterance.

Evaluation of SF and IC on more complex dataset. Existing neural approaches typically evaluated on *single-intent* utterance, however in a real-world scenario users may indicate *multiple-intent* in an utterance e.g. "*Show me all flights from Atlanta to London and get the cost*" (Gangadharaiah and Narayanaswamy, 2019) or even expressing multiple sentences in one single turn. While most datasets for slot filling and intent classification are *single-turn* utterance, there are some recent multi-turn datasets that provide slot annotation on the token-level, namely the RESTAURANT-8K, TaskMaster-1 and 2 (Byrne et al., 2019), and Frame (Asri et al., 2017) datasets. The subset of Schema Guided Dialogue (SGD) dataset (Rastogi et al., 2020) used in DTSC-8 is also annotated with slots in the token-level and covers 16 domains. In addition to that, the TOP dataset (Gupta et al., 2018) introduces datasets annotated with *hierarchical* representation and MTOP dataset (Li et al., 2020) provides both flat and hierarchical representation on 6 languages across 11 domains.

7 Conclusion

We have surveyed recent neural-based models applied to SF and IC in the context of task-oriented dialogue systems. We examined three approaches, i.e. *independent*, *joint*, and *transfer learning based* models. Joint models exploiting the relation between SF and IC simultaneously shown relatively better performance than independent models. Empirical results have shown that most joint models nearly "solve" widely used datasets, ATIS and SNIPS, given *sufficient in-domain training data*. Nevertheless, there are still several challenges related to SF and IC, especially improving the scalability of the model to new domains and languages when limited labeled data are available.

References

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 Kristiina Jokinen, Manfred Stede, David DeVault, and Annie Louis, editors, *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, August 15-17, 2017*, pages 207–219. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ankur Bapna, Gökhan Tür, Dilek Hakkani-Tür, and Larry P. Heck. 2017. Towards zero-shot frame semantic parsing for domain scaling. In Francisco Lacerda, editor, *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 2476–2480. ISCA.

Frédéric Béchet and Christian Raymond. 2018. Is ATIS too shallow to go deeper for benchmarking spoken language understanding models? In B. Yegnanarayana, editor, *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 3449–3453. ISCA.

Jerome R. Bellegarda. 2013. Large-scale personal assistant technology deployment: the siri experience. In Frédéric Bimbot, Christophe Cerisara, Cécile Fougeron, Guillaume Gravier, Lori Lamel, François Pellegrino, and Pascal Perrier, editors, *INTERSPEECH*, pages 2029–2033. ISCA.

Valentina Bellomaria, Giuseppe Castellucci, Andrea Favalli, and Raniero Romagnoli. 2019. Almawave-slu: A new dataset for SLU in italian. In Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Hélène Bonneau-Maynard, Sophie Rosset, Christelle Ayache, Anne Kuhn, and Djamel Mostefa. 2005. Semantic annotation of the french media dialog corpus. In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, pages 3457–3460. ISCA.

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

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4515–4524. Association for Computational Linguistics.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *ArXiv*, abs/1902.10909.

Eunah Cho, He Xie, John P. Lalor, Varun Kumar, and William M. Campbell. 2019. Efficient semi-supervised learning for natural language understanding by optimizing diversity. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, pages 1077–1084. IEEE.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 160–167. ACM.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *ArXiv*, abs/1805.10190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Quynh Ngoc Thi Do and Judith Gaspers. 2019. Cross-lingual transfer learning with data selection for large-scale spoken language understanding. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1455–1460. Association for Computational Linguistics.

Haihong E, Peiqing Niu, Zhongfu Chen, and Meina Song. 2019. A novel bi-directional interrelated model for joint intent detection and slot filling. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5467–5471. Association for Computational Linguistics.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Rashmi Gangadharaiah and Balakrishnan Narayanaswamy. 2019. Joint multiple intent detection and slot labeling for goal-oriented dialog. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 564–569. Association for Computational Linguistics.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 753–757. Association for Computational Linguistics.

Anuj Kumar Goyal, Angeliki Metallinou, and Spyros Matsoukas. 2018. Fast and scalable expansion of natural language understanding functionality for intelligent agents. In Srinivas Bangalore, Jennifer Chu-Carroll, and Yunyao Li, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 3 (Industry Papers)*, pages 145–152. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference- 6: A brief history. In *16th International Conference on Computational Linguistics, Proceedings of the Conference, COLING 1996, Center for Sprogteknologi, Copenhagen, Denmark, August 5-9, 1996*, pages 466–471.

Marco Guerini, Simone Magnolini, Vevake Balaraman, and Bernardo Magnini. 2018. Toward zero-shot entity recognition in task-oriented conversational agents. In Kazunori Komatani, Diane J. Litman, Kai Yu, Lawrence Cavedon, Mikio Nakano, and Alex Papangelis, editors, *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, Melbourne, Australia, July 12-14, 2018*, pages 317–326. Association for Computational Linguistics.

Daniel Guo, Gökhan Tür, Wen-tau Yih, and Geoffrey Zweig. 2014. Joint semantic utterance classification and slot filling with recursive neural networks. In *2014 IEEE Spoken Language Technology Workshop, SLT 2014, South Lake Tahoe, NV, USA, December 7-10, 2014*, pages 554–559. IEEE.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2787–2792. Association for Computational Linguistics.

Dilek Hakkani-Tür, Gökhan Tür, Asli Çelikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM. In Nelson Morgan, editor, *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 715–719. ISCA.

Tianxing He, Juefu Liu, Kyunghyun Cho, Myle Ott, Bing Liu, James Glass, and Fuchun Peng. 2019. Analyzing the forgetting problem in the pretrain-finetuning of dialogue response models. *arXiv: Computation and Language*.

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

Matthew Henderson, Blaise Thomson, and Steve Young. 2014. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 292–299.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1234–1245. Association for Computational Linguistics.

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1381–1393. Association for Computational Linguistics.

Aaron Jaech, Larry P. Heck, and Mari Ostendorf. 2016. Domain adaptation of recurrent neural networks for natural language understanding. In Nelson Morgan, editor, *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 690–694. ISCA.

Minwoo Jeong and Gary Geunbae Lee. 2008. Triangular-chain conditional random fields. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(7):1287–1302.

Rahul Jha, Alex Marin, Suvamsh Shivaprasad, and Imed Zitouni. 2018. Bag of experts architectures for model reuse in conversational language understanding. In Srinivas Bangalore, Jennifer Chu-Carroll, and Yunyao Li, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 3 (Industry Papers)*, pages 153–161. Association for Computational Linguistics.

Michael I Jordan. 1997. Serial order: A parallel distributed processing approach. In *Advances in psychology*, volume 121, pages 471–495. Elsevier.

Young-Bum Kim, Karl Stratos, and Dongchan Kim. 2017. Domain attention with an ensemble of experts. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 643–653. Association for Computational Linguistics.

Hwa-Yeon Kim, Yoon-Hyung Roh, and Young-Kil Kim. 2019. Data augmentation by data noising for open-vocabulary slots in spoken language understanding. In Sudipta Kar, Farah Nadeem, Laura Burdick, Greg Durrett, and Na-Rae Han, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 3-5, 2019, Student Research Workshop*, pages 97–102. Association for Computational Linguistics.

Gakuto Kurata, Bing Xiang, Bowen Zhou, and Mo Yu. 2016. Leveraging sentence-level information with encoder LSTM for semantic slot filling. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2077–2083. The Association for Computational Linguistics.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Sungjin Lee and Rahul Jha. 2019. Zero-shot adaptive transfer for conversational language understanding. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6642–6649. AAAI Press.

Changliang Li, Liang Li, and Ji Qi. 2018. A self-attentive model with gate mechanism for spoken language understanding. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3824–3833. Association for Computational Linguistics.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2020. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. *CoRR*, abs/2008.09335.

Bing Liu and Ian Lane. 2015. Recurrent neural network structured output prediction for spoken language understanding.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. In Nelson Morgan, editor, *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 685–689. ISCA.

Jingjing Liu, Panupong Pasupat, Scott Cyphers, and James R. Glass. 2013. Asgard: A portable architecture for multilingual dialogue systems. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pages 8386–8390. IEEE.

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Zero-shot cross-lingual dialogue systems with transferable latent variables. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1297–1303. Association for Computational Linguistics.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020a. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. 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 8433–8440. AAAI Press.

Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung. 2020b. Coach: A coarse-to-fine approach for cross-domain slot filling. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 19–25. Association for Computational Linguistics.

Samuel Louvan and Bernardo Magnini. 2019. Leveraging non-conversational tasks for low resource slot filling: Does it help? In Satoshi Nakamura, Milica Gasic, Ingrid Zuckerman, Gabriel Skantze, Mikio Nakano, Alexandros Papangelis, Stefan Ultes, and Koichiro Yoshino, editors, *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, pages 85–91. Association for Computational Linguistics.

Grégoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio. 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In Frédéric Bimbot, Christophe Cerisara, Cécile Fougeron, Guillaume Gravier, Lori Lamel, François Pellegrino, and Pascal Perrier, editors, *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, pages 3771–3775. ISCA.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Z. Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, and Geoffrey Zweig. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23:530–539.

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura, editors, *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048. ISCA.

Tomas Mikolov, Stefan Kombrink, Anoop Deoras, Lukás Burget, and Jan Cernocky. 2011. Rnnlm - recurrent neural network language modeling toolkit.

Nikola Mrksic, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasic, Pei-hao Su, David Vandyke, Tsung-Hsien Wen, and Steve J. Young. 2015. Multi-domain dialog state tracking using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 794–799. The Association for Computer Linguistics.

Jingcheng Niu and Gerald Penn. 2019. Rationally reappraising atis-based dialogue systems. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5503–5507. Association for Computational Linguistics.

Baolin Peng, Chenguang Zhu, Michael Zeng, and Jianfeng Gao. 2020. Data augmentation for spoken language understanding via pretrained models. *CoRR*, abs/2004.13952.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2078–2087. Association for Computational Linguistics.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3853–3860. ijcai.org.

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.

Suman V. Ravuri and Andreas Stolcke. 2015. Recurrent neural network and LSTM models for lexical utterance classification. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 135–139. ISCA.

Christian Raymond and Giuseppe Riccardi. 2007. Generative and discriminative algorithms for spoken language understanding. In *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*, pages 1605–1608. ISCA.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147. ACL.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3795–3805. Association for Computational Linguistics.

Darsh Shah, Raghav Gupta, Amir Fayazi, and Dilek Hakkani-Tur. 2019. Robust zero-shot cross-domain slot filling with example values. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5484–5490, Florence, Italy, July. Association for Computational Linguistics.

Aditya Siddhant, Anuj Kumar Goyal, and Angeliki Metallinou. 2019. Unsupervised transfer learning for spoken language understanding in intelligent agents. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 4959–4966. AAAI Press.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Raymond Hendy Susanto and Wei Lu. 2017. Neural architectures for multilingual semantic parsing. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 38–44. Association for Computational Linguistics.

Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Shyam Upadhyay, Manaal Faruqui, Gökhan Tür, Dilek Z. Hakkani-Tür, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Ngoc Thang Vu, Pankaj Gupta, Heike Adel, and Hinrich Schütze. 2016. Bi-directional recurrent neural network with ranking loss for spoken language understanding. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6060–6064. IEEE.

Ye-Yi Wang, Li Deng, and Alex Acero. 2005. Spoken language understanding. *IEEE Signal Processing Magazine*, 22(5):16–31.

Yu Wang, Yilin Shen, and Hongxia Jin. 2018. A bi-model based RNN semantic frame parsing model for intent detection and slot filling. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 309–314. Association for Computational Linguistics.

Puyang Xu and Ruhi Sarikaya. 2013. Convolutional neural network based triangular crf for joint intent detection and slot filling. In *2013 ieee workshop on automatic speech recognition and understanding*, pages 78–83. IEEE.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. *CoRR*, abs/2004.14353.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Kaisheng Yao, Geoffrey Zweig, Mei-Yuh Hwang, Yangyang Shi, and Dong Yu. 2013. Recurrent neural networks for language understanding. In Frédéric Bimbot, Christophe Cerisara, Cécile Fougeron, Guillaume Gravier, Lori Lamel, François Pellegrino, and Pascal Perrier, editors, *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, pages 2524–2528. ISCA.

Kaisheng Yao, Baolin Peng, Yu Zhang, Dong Yu, Geoffrey Zweig, and Yangyang Shi. 2014a. Spoken language understanding using long short-term memory neural networks. *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 189–194.

Kaisheng Yao, Baolin Peng, Geoffrey Zweig, Dong Yu, Xiaolong Li, and Feng Gao. 2014b. Recurrent conditional random field for language understanding. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4077–4081.

Dani Yogatama, Chris Dyer, Wang Ling, and Phil Blunsom. 2017. Generative and discriminative text classification with recurrent neural networks. *ArXiv*, abs/1703.01898.

Kang Min Yoo, Youhyun Shin, and Sang-goo Lee. 2019. Data augmentation for spoken language understanding via joint variational generation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7402–7409. AAAI Press.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In Subbarao Kambhampati, editor, *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2993–2999. IJCAI/AAAI Press.

Zhichang Zhang, Zhenwen Zhang, Haoyuan Chen, and Zhiman Zhang. 2019. A joint learning framework with bert for spoken language understanding. *IEEE Access*, 7:168849–168858.

Zijian Zhao, Su Zhu, and Kai Yu. 2019. Data augmentation with atomic templates for spoken language understanding. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3635–3641. Association for Computational Linguistics.

Appendix A. SF and IC Dataset

Dataset	**Language**	**# intent**	**# slot**	**# sentences train/dev/test**
ATIS	English	18	83	4,478 / 500 / 893
MEDIA	French	-	68	12,908/1,259/3,005
SNIPS-NLU	English	7	39	13,084 / 700 / 700
	Italian	7	39	5,742 /700 / 700
Facebook	English	12	11	30,521 / 4,181 / 8,621
Multilingual	Thai	12	11	3,617 / 1,983 / 3,043
	Spanish	12	11	2,156 / 1,235 / 1,692
MIT Restaurant	English	-	8	6,128 / 1,532 / 1,521
MIT Movie	English	-	12	7,820 / 1,955 / 2,443

Table 4: Single-turn datasets statistics.