Intent Detection with WikiHow

Li Zhang **Qing Lyu** **Chris Callison-Burch**
University of Pennsylvania
{zharry,lyuqing,ccb}@seas.upenn.edu

Abstract

Modern task-oriented dialog systems need to reliably understand users' intents. Intent detection is even more challenging when moving to new domains or new languages, since there is little annotated data. To address this challenge, we present a suite of pretrained intent detection models which can predict a broad range of intended goals from many actions because they are trained on wikiHow, a comprehensive instructional website. Our models achieve state-of-the-art results on the Snips dataset, the Schema-Guided Dialogue dataset, and all 3 languages of the Facebook multilingual dialog datasets. Our models also demonstrate strong zero- and few-shot performance, reaching over 75% accuracy using only 100 training examples in all datasets.[1]

1 Introduction

Task-oriented dialog systems like Apple's Siri, Amazon Alexa, and Google Assistant have become pervasive in smartphones and smart speakers. To support a wide range of functions, dialog systems must be able to map a user's natural language instruction onto the desired skill or API. Performing this mapping is called intent detection.

Intent detection is usually formulated as a sentence classification task. Given an utterance (e.g. "wake me up at 8"), a system needs to predict its intent (e.g. "Set an Alarm"). Most modern approaches use neural networks to jointly model intent detection and slot filling (Xu and Sarikaya, 2013; Liu and Lane, 2016; Goo et al., 2018; Zhang et al., 2019). In response to a rapidly growing range of services, more attention has been given to zero-shot intent detection (Ferreira et al., 2015a,b; Yazdani and Henderson, 2015; Chen et al., 2016; Kumar et al., 2017; Gangadharaiah and Narayanaswamy, 2019). While most existing research on intent detection proposed novel model architectures, few have attempted data augmentation. One such work (Hu et al., 2009) showed that models can learn much knowledge that is important for intent detection from massive online resources such as Wikipedia.

We propose a pretraining task based on wikiHow, a comprehensive instructional website with over 110,000 professionally edited articles. Their topics span from common sense such as "How to Download Music" to more niche tasks like "How to Crochet a Teddy Bear." We observe that the header of each step in a wikiHow article describes an action and can be approximated as an utterance, while the title describes a goal and can be seen as an intent. For example, "find good gas prices" in the article "How to Save Money on Gas" is similar to the utterance "where can I find cheap gas?" with the intent "Save Money on Gas." Hence, we introduce a dataset based on wikiHow, where a model predicts the goal of an action given some candidates. Although most of wikiHow's domains are far beyond the scope of any present dialog system, models pretrained on our dataset would be robust to emerging services and scenarios. Also, as wikiHow is available in 18 languages, our pretraining task can be readily extended to multilingual settings.

Using our pretraining task, we fine-tune transformer language models, achieving state-of-the-art results on the intent detection task of the Snips dataset (Coucke et al., 2018), the Schema-Guided Dialog (SGD) dataset (Rastogi et al., 2019), and all 3 languages (English, Spanish, and Thai) of the Facebook multilingual dialog datasets (Schuster et al., 2019), with statistically significant improvements. As our accuracy is close to 100% on all these datasets, we further experiment with zero- or few-shot settings. Our models achieve over 70% accuracy with no in-domain training data on Snips

[1]The data and models are available at https://github.com/zharry29/wikihow-intent.

Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 328–333

and SGD, and over 75% with only 100 training examples on all datasets. This highlights our models' ability to quickly adapt to new utterances and intents in unseen domains.

2 WikiHow Pretraining Task

2.1 Corpus

We crawl the wikiHow website in English, Spanish, and Thai (the languages were chosen to match those in the Facebook multilingual dialog datasets). We define the **goal** of each artcle as its title stripped of the prefix "How to" (and its equivalent in other languages). We extract a set of **steps** for each article, by taking the bolded header of each paragraph.

2.2 WikiHow Pretraining Dataset

A wikiHow article's goal can approximate an intent, and each step in it can approximate an associated utterance. We formulate the pretraining task as a 4-choose-1 multiple choice format: given a step, the model infers the correct goal among 4 candidates. For example, given the step "let check-in agents and flight attendants know if it's a special occasion" and the candidate goals:

A. Get Upgraded to Business Class
B. Change a Flight Reservation
C. Check Flight Reservations
D. Use a Discount Airline Broker

the correct goal would be A. This is similar to intent detection, where a system is given a user utterance and then must select a supported intent.

We create intent detection pretraining data using goal-step pairs from each wikiHow article. Each article contributes at least one positive goal-step pair. However, it is challenging to sample negative candidate goals for a given step. There are two reasons for this. First, random sampling of goals correctly results in true negatives, but they tend to be so distant from the positive goal that the classification task becomes trivial and the model does not learn sufficiently. Second, if we sample goals that are similar to the positive goal, then they might not be true negatives, since there are many steps in wikiHow often with overlapping goals. To sample high-quality negative training instances, we start with the correct goal and search in its article's "related articles" section for an article whose title has the least lexical overlap with the current goal. We recursively do this until we have enough candidates. Empirically, examples created this way are mostly clean, with an example shown above. We select one positive goal-step pair from each article by picking its longest step. In total, our wikiHow pretraining datasets have 107,298 English examples, 64,803 Spanish examples, and 6,342 Thai examples.

3 Experiments

We fine-tune a suite of off-the-shelf language models pretrained on our wikiHow data, and evaluate them on 3 major intent detection benchmarks.

3.1 Models

We fine-tune a pretrained RoBERTa model (Liu et al., 2019) for the English datasets and a pretrained XLM-RoBERTa model (Conneau et al., 2019) for the multilingual datasets. We cast the instances of the intent detection datasets into a multiple-choice format, where the utterance is the input and the full set of intents are the possible candidates, consistent with our wikiHow pretraining task. For each model, we append a linear classification layer with cross-entropy loss to calculate a likelihood for each candidate, and output the candidate with the maximum likelihood.

For each intent detection dataset in any language, we consider the following settings:

+in-domain (+ID): a model is only trained on the dataset's in-domain training data;

+wikiHow +in-domain (+WH+ID): a model is first trained on our wikiHow data in the corresponding language, and then trained on the dataset's in-domain training data;

+wikiHow zero-shot (+WH 0-shot): a model is trained only on our wikiHow data in the corresponding language, and then applied directly to the dataset's evaluation data.

For non-English languages, the corresponding wikiHow data might suffer from smaller sizes and lower quality. Hence, we additionally consider the following cross-lingual transfer settings for non-English datasets:

+en wikiHow +in-domain (+enWH+ID), a model is trained on wikiHow data in English, before it is trained on the dataset's in-domain training data;

+en wikiHow zero-shot (+enWH 0-shot), a model is trained on wikiHow data in English, before it is directly applied to the dataset's evaluation data.

3.2 Datasets

We consider the 3 following benchmarks:

The Snips dataset (Coucke et al., 2018) is a single-turn English dataset. It is one of the most cited dialog benchmarks in recent years, containing

	Training Size	Valid. Size	Test Size	Num. Intents
Snips	2,100	700	N/A	7
SGD	163,197	24,320	42,922	4
FB-en	30,521	4,181	8,621	12
FB-es	3,617	1,983	3,043	12
FB-th	2,156	1,235	1,692	12

Table 1: Statistics of the dialog benchmark datasets.

utterances collected from the Snips personal voice assistant. While its full training data has 13,784 examples, we find that our models only need its smaller training split consisting of 2,100 examples to achieve high performance. Since Snips does not provide test sets, we use the validation set for testing and the full training set for validation. Snips involves 7 intents, including *Add to Playlist*, *Rate Book*, *Book Restaurant*, *Get Weather*, *Play Music*, *Search Creative Work*, and *Search Screening Event*. Some example utterances include "Play the newest melody on Last Fm by Eddie Vinson," "Find the movie schedule in the area," etc.

The Schema-Guided Dialogue dataset (SGD) (Rastogi et al., 2019) is a multi-turn English dataset. It is the largest dialog corpus to date spanning dozens of domains and services, used in the DSTC8 challenge (Rastogi et al., 2020) with dozens of team submissions. Schemas are provided with at most 4 intents per dialog turn. Examples of these intents include *Buy Movie Tickets for a Particular show*, *Make a Reservation with the Therapist*, *Book an Appointment at a Hair Stylist*, *Browse attractions in a given city*, etc. At each turn, we use the last 3 utterances as input. An example: "That sounds fun. What other attractions do you recommend? There is a famous place of worship called Akshardham."

The Facebook multilingual datasets (FB-en/es/th) (Schuster et al., 2019) is a single-turn multilingual dataset. It is the only multilingual dialog dataset to the best of our knowledge, containing utterances annotated with intents and slots in English (en), Spanish (es), and Thai (th). It involves 12 intents, including *Set Reminder*, *Check Sunrise*, *Show Alarms*, *Check Sunset*, *Cancel Reminder*, *Show Reminders*, *Check Time Left on Alarm*, *Modify Alarm*, *Cancel Alarm*, *Find Weather*, *Set Alarm*, and *Snooze Alarm*. Some example utterances are "Is my alarm set for 10 am today?" "Colocar una alarma para mañana a las 3 am," ฉันมีนาฬิกาปลุกสำหรับตอนเช้ามั๊ย etc.

	Snips	SGD	FB-en
(Ren and Xue, 2020)	.993	N/A	.993
(Ma et al., 2019)	N/A	.948	N/A
+in-domain (+ID)	.990	.942	.993
(ours) +WH+ID	**.994**	**.951**†	**.995**†
(ours) +WH 0-shot	.713	.787	.445
Chance	.143	.250	.083

Table 2: The accuracy of intent detection on English datasets using RoBERTa. State-of-the-art performances are in bold; † indicates statistically significant improvement from the previous state-of-the-art.

	FB-en	FB-es	FB-th
(Ren and Xue, 2020)	.993	N/A	N/A
(Zhang et al., 2019)	N/A	.978	.967
+in-domain (+ID)	.993	.986	.962
(ours) +WH+ID	**.995**	.988	.971
(ours) +enWH+ID	**.995**	**.990**†	**.976**†
(ours) +WH 0-shot	.416	.129	.119
(ours) +enWH 0-shot	.416	.288	.124
Chance	.083	.083	.083

Table 3: The accuracy of intent detection on multilingual datasets using XLM-RoBERTa.

Statistics of the datasets are shown in Table 1.

3.3 Baselines

We compare our models with the previous state-of-the-art results of each dataset:

• Ren and Xue (2020) proposed a Siamese neural network with triplet loss, achieving state-of-the-art results on Snips and FB-en;

• Zhang et al. (2019) used multi-task learning to jointly learn intent detection and slot filling, achieving state-of-the-art results on FB-es and FB-th;

• Ma et al. (2019) augmented the data via back-translation to and from Chinese, achieving state-of-the-art results on SGD.

3.4 Modelling Details

After experimenting with base and large models, we use RoBERTa-large for the English datasets and XLM-RoBERTa-base for the multilingual dataset for best performances. All our models are implemented using the HuggingFace Transformer library[2].

We tune our model hyperparameters on the validation sets of the datasets we experiment with. However, in all cases, we use a unified setting

[2] https://github.com/huggingface/transformers

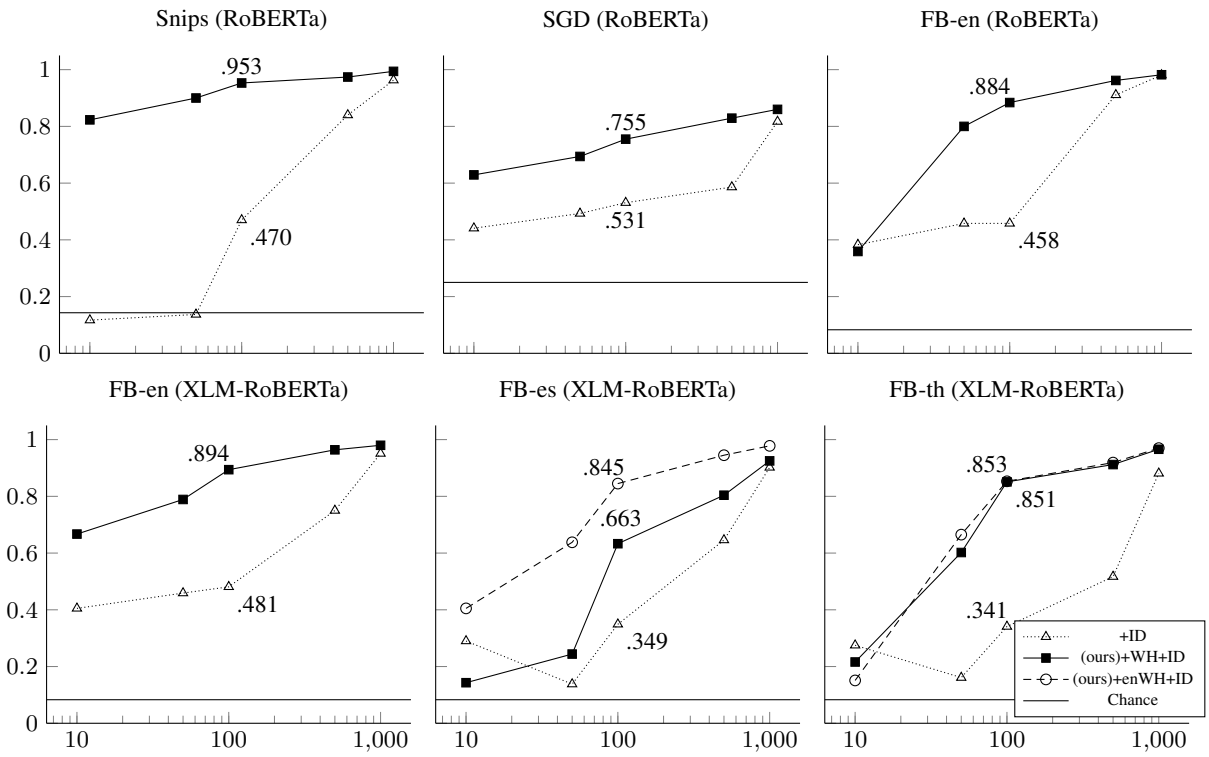

Figure 1: Learning curves of models in low-resource settings. The vertical axis is the accuracy of intent detection, while the horizontal axis is the number of in-domain training examples of each task, distorted to log-scale.

which empirically performs well, using the Adam optimizer (Kingma and Ba, 2014) with an epsilon of $1e^{-8}$, a learning rate of $5e^{-6}$, maximum sequence length of 80 and 3 epochs. We variate the batch size from 2 to 16 according to the number of candidates in the multiple-choice task, to avoid running out of memory. We save the model every 1,000 training steps, and choose the model with the highest validation performance to be evaluated on the test set.

We run our experiments on an NVIDIA GeForce RTX 2080 Ti GPU, with half-precision floating point format (FP16) with O1 optimization. Each epoch takes up to 90 minutes in the most resource intensive setting, i.e. running a RoBERTa-large on around 100,000 training examples of our wikiHow pretraining dataset.

3.5 Results

The performance of RoBERTa on the English datasets (Snips, SGD, and FB-en) are shown in Table 2. We repeat each experiment 20 times, report the mean accuracy, and calculate its p-value against the previous state-of-the-art result, using a one-sample and one-tailed t-test with a significance level of 0.05. Our models achieve state-of-the-art results using the available in-domain training data. Moreover, our wikiHow data enables our models to demonstrate strong performances in zero-shot settings with no in-domain training data, implying our models' strong potential to adapt to new domains.

The performance of XLM-RoBERTa on the multilingual datasets (FB-en, FB-es, and FB-th) are shown in Table 3. Our models achieve state-of-the-art results on all 3 languages. While our wikiHow data in Spanish and Thai does improve models' performances, its effect is less salient than the English wikiHow data.

Our experiments above focus on settings where all available in-domain training data are used. However, modern task-oriented dialog systems must rapidly adapt to burgeoning services (e.g. Alexa Skills) in different languages, where little training data are available. To simulate low-resource settings, we repeat the experiments with exponentially increasing number of training examples up to 1,000. We consider the models trained only on in-domain data (+ID), those first pretrained on our wikiHow data in corresponding languages (+WH+ID), and those first pretrained on our English wikiHow data (+enWH+ID) for FB-es and FB-th.

The learning curves of each dataset are shown in Figure 1. Though the vanilla transformers models (+ID) achieve close to state-of-the-art performance with access to the full training data (see Table 2

and 3), they struggle in the low-resource settings. When given up to 100 in-domain training examples, their accuracies are less than 50% on most datasets. In contrast, our models pretrained on our wikiHow data (+WH+ID) can reach over 75% accuracy given only 100 training examples on all datasets.

4 Discussion and Future Work

As our model performances exceed 99% on Snips and FB-en, the concern arises that these intent detection datasets are "solved". We address this by performing error analysis and providing future outlooks for intent detection.

4.1 Error Analysis

Our model misclassifies 7 instances in the Snips test set. Among them, 6 utterances include proper nouns on which intent classification is contingent. For example, the utterance "please open Zvooq" assumes the knowledge that Zvooq is a streaming service, and its labelled intent is "Play Music."

Our model misclassifies 43 instances in the FB-en test set. Among them, 10 has incorrect labels: e.g. the labelled intent of "have alarm go off at 5 pm" is "Show Alarms," while our model prediction "Set Alarm" is in fact correct. 28 are ambiguous: e.g. the labelled intent of "repeat alarm every weekday" is "Set Alarm," whereas that of "add an alarm for 2:45 on every Monday" is "Modify Alarm." We only find 1 example an interesting edge case: the gold intent of "remind me if there will be a rain forecast tomorrow" is "Find Weather," while our model incorrectly chooses "Set Reminder."

By performing manual error analyses on our model predictions, we observe that most misclassified examples involve ambiguous wordings, wrong labels, or obscure proper nouns. Our observations imply that Snips and FB-en might be too easy to effectively evaluate future models.

4.2 Open-Domain Intent Detection

State-of-the-art models now achieve greater than 99% percent accuracy on standard benchmarks for intent detection. However, intent detection is far from being solved. The standard benchmarks only have a dozen intents, but future dialog systems will need to support many more functions with intents from a wide range of domains. To demonstrate that our pretrained models can adapt to unseen, open-domain intents, we hold out 5,000 steps (as utterances) with their corresponding goals (as intents) from our wikiHow dataset as a proxy of an intent detection dataset with more than 100,000 possible intents (all goals in wikiHow).

For each step, we sample 100 goals with the highest embedding similarity to the correct goal, as most other goals are irrelevant. We then rank them for the likelihood that the step helps achieve them. Our RoBERTa model achieves a mean reciprocal rank of 0.462 and a 36% accuracy of ranking the correct goal first. As a qualitative example, given the step "find the order that you want to cancel," the top 3 ranked steps are "Cancel an Order on eBay", "Cancel an Online Order", "Cancel an Order on Amazon." This hints that our pretrained models' can work with a much wider range of intents than those in current benchmarks, and suggests that future intent detection research should focus on open domains, especially those with little data.

5 Conclusion

By pretraining language models on wikiHow, we attain state-of-the-art results in 5 major intent detection datasets spanning 3 languages. The wide-ranging domains and languages of our pretraining resource enable our models to excel with few labelled examples in multilingual settings, and suggest open-domain intent detection is now feasible.

Acknowledgments

This research is based upon work supported in part by the DARPA KAIROS Program (contract FA8750-19-2-1004), the DARPA LwLL Program (contract FA8750-19-2-0201), and the IARPA BETTER Program (contract 2019-19051600004). Approved for Public Release, Distribution Unlimited. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, IARPA, or the U.S. Government.

We thank the anonymous reviewers for their valuable feedback.

References

Yun-Nung Chen, Dilek Hakkani-Tür, and Xiaodong He. 2016. Zero-shot learning of intent embeddings for expansion by convolutional deep structured semantic models. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6045–6049. IEEE.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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

Emmanuel Ferreira, Bassam Jabaian, and Fabrice Lefèvre. 2015a. Zero-shot semantic parser for spoken language understanding. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Emmanuel Ferreira, Bassam Jabaian, and Fabrice Lefèvre. 2015b. Online adaptative zero-shot learning spoken language understanding using word-embedding.

Rashmi Gangadharaiah and Balakrishnan Narayanaswamy. 2019. Joint multiple intent detection and slot labeling for goal-oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 564–569, Minneapolis, Minnesota. 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 *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757, New Orleans, Louisiana. Association for Computational Linguistics.

Jian Hu, Gang Wang, Fred Lochovsky, Jian-tao Sun, and Zheng Chen. 2009. Understanding user's query intent with wikipedia. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, page 471–480, New York, NY, USA. Association for Computing Machinery.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Anjishnu Kumar, Pavankumar Reddy Muddireddy, Markus Dreyer, and Björn Hoffmeister. 2017. Zero-shot learning across heterogeneous overlapping domains.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling.

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

Yue Ma, Zengfeng Zeng, Dawei Zhu, Xuan Li, Yiying Yang, Xiaoyuan Yao, Kaijie Zhou, and Jianping Shen. 2019. An end-to-end dialogue state tracking system with machine reading comprehension and wide & deep classification.

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

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Schema-guided dialogue state tracking task at dstc8.

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

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

P. Xu and R. 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.

Majid Yazdani and James Henderson. 2015. A model of zero-shot learning of spoken language understanding. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 244–249, Lisbon, Portugal. Association for Computational Linguistics.

Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip Yu. 2019. Joint slot filling and intent detection via capsule neural networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5259–5267, Florence, Italy. Association for Computational Linguistics.

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