Locale-agnostic Universal Domain Classification Model in Spoken Language Understanding

Jihwan Lee
Amazon Alexa AI
jihwl@amazon.com

Ruhi Sarikaya
Amazon Alexa AI
rsarikay@amazon.com

Young-Bum Kim
Amazon Alexa AI
youngbum@amazon.com

Abstract

In this paper, we introduce an approach for leveraging available data across multiple locales sharing the same language to 1) improve domain classification model accuracy in Spoken Language Understanding and user experience even if new locales do not have sufficient data and 2) reduce the cost of scaling the domain classifier to a large number of locales. We propose a locale-agnostic universal domain classification model based on selective multi-task learning that learns a joint representation of an utterance over locales with different sets of domains and allows locales to share knowledge selectively depending on the domains. The experimental results demonstrate the effectiveness of our approach on domain classification task in the scenario of multiple locales with imbalanced data and disparate domain sets. The proposed approach outperforms other baselines models especially when classifying locale-specific domains and also low-resourced domains.

1 Introduction

Recent success of intelligent personal digital assistants (IPDA) such as Amazon Alexa, Google Assistant, Apple Siri, Microsoft Cortana (Sarikaya, 2017; Sarikaya et al., 2016) in USA has led to their expansion to multiple locales and languages. Some of those virtual assistant systems have been released in the United States (US), the United Kingdom (GB), Canada (CA), India (IN), and so on. Such expansion typically leads to building a separate domain classification model for each new locale, and it brings two challenging issues: 1) having a separate model per locale becomes a bottleneck for rapid scaling of virtual assistant due to the resource and maintenance costs that grow linearly with the number of locales, and 2) new locales typically comes without much training data and cannot take full advantage of useful data available in other mature locales to achieve the high model accuracy.

In this study, we propose a new approach that reduces the cost of scaling natural language understanding to a large number of locales, given the sufficient amount of data in one of the locales of that language, while achieving high domain classification accuracy over all locales. The approach is based on a multi-task learning framework that aims to share available data to learn a joint representation, and we introduce a way to selectively share knowledge across locales while considering locale-specificity in the joint learning. Multi-task learning has been widely used to tackle the problem of low-resource tasks or leveraging data between correlated targets (Liu et al., 2017; Ruder and Plank, 2018; Augenstein et al., 2018; Peters et al., 2017; Kim et al., 2017b), but none of them consider locale-specificity when sharing knowledge to learn a joint representation.

We evaluate our proposed approach on the real-world utterance data spoken by customers to an intelligent personal digital assistant across different locales. The experimental results empirically demonstrate that the proposed universal model scales to multiple locales, while achieving higher domain classification accuracy compared to competing locale-unified models as well as per-locale separate models. The proposed model named universal model is able to successfully predict domains for locale-specific utterances while sharing common knowledge across locales without sacrificing the accuracy of predicting locale-independent domains.

The paper is organized as follows. In Section 2, we discuss several design considerations that motivate our model design. In Section 3, we define the problem of domain classification with multiple locales that have different domain sets, and then introduce a novel universal domain classifi-

Proceedings of NAACL-HLT 2019, pages 9–15
Minneapolis, Minnesota, June 2 - June 7, 2019.

cation model with several technical details. We present our experimental observations over different approaches on the Amazon Alexa dataset in Section 4. Finally, we conclude the paper in Section 5.

2 Motivations

2.1 Locale/Domain-Maturity

Let the term *maturity* be defined by how long it has been since a service or model was deployed in a locale and/or how much data have been collected. Every locale has different degrees of maturity. That is, while some locales have spent time long enough to collect sufficient data to train models, others may suffer from the lack of data (see more details of data statistics in Section 4). In addition to that, domains that are commonly available in multiple locales have different levels of maturity for each locale. Those two dimensions of maturity are not always aligned with each other. In other words, there could exist domains that have more data in immature locales than in mature locales, depending on targeted users, regional properties of domains, and so forth.

2.2 Locale-Specificity

When an SLU service is deployed in multiple locales, each of the locales has its own domain set and there can exist overlapping domains between locales. Such domains may share the same schema including intents and slots and thus they should be able to handle the same patterns of utterances regardless of locales. It allows locales to share the knowledge of common domains with each other, which eventually helps immature locales to overcome the lack of data. A special case that needs to be carefully considered is that a domain could be locale-specific. Even though a domain is common across different locales, it may be defined with different intents/slots. For example, the domain `OpenTable`, which is capable of restaurant reservation, is available in both US and GB, but the slot values including restaurant names are totally different between the two locales. That is, the utterance *"Make a reservation for The Fox Club London"* can be handled by `OpenTable` in GB locale, but probably not in US locale, because the restaurant *The Fox Club London* is located in London. If we have different locales share the same utterance patterns between them even for such locale-specific domains, then it will cause confusion on the models. We thus identify locale-specific domains in advance of model training and do not allow input utterances of such domain to be shared by different locales. We need to handle domains in a similar way that are available only in a particular locale.

3 Universal Model

In this section, we describe our proposed model illustrated in Figure 1 in detail. Suppose that given k locales, $\{l_i | i = 1, 2, \ldots, k\}$, each locale l_i is associated with its own domain set $D_i = \{d_{ij} | j = 1, 2, \ldots, |D_i|\}$. There could exist overlapping domains between locales and some of the overlapping domains may share exactly the same intents/slots while others may have different intents/slots across locales. The main task is that given an input utterance from locale l_i the model should be able to correctly classify the utterance into a domain $d_{ij} \in D_i$ that can best handle the utterance. Here we assume that all locales use the same language, English, but have different domain sets. Our deep neural model, as a proposed solution to the task, is comprised of two layers. The first layer includes a BiLSTM shared encoder and k BiLSTM locale-specific encoders. The second layer consists of a set of k locale-specific prediction layers.

3.1 Shared and Locale-specific Encoders

Given an input utterance that forms a word sequence, an encoder makes a vector representation of the entire utterance by using word embeddings for English language in general. We use Bidirectional LSTM (BiLSTM) to encode an input utterance and consider it to be a mapping function $\mathcal{F}$ that consumes a sequence of word embeddings and then produces an embedding vector given by concatenating the outputs of the ends of the word sequences from the forward LSTM and the backward LSTM. While different locales share common domains and utterances, each of them also should be able to learn certain patterns observed from domains available only in the locale. In other words, there exist both global and local patterns in the entire domain set. In order to effectively capture both patterns and avoid confusion between locales, we use a shared encoder $\mathcal{F}_s$ and multiple locale-specific encoders $\mathcal{F}_{l_i}$ for $\forall i = 1, 2, \ldots, k$, each of which corresponds to a particular locale l_i, as similarly adopted in (Kim et al., 2017a, 2016c). While the shared encoder $\mathcal{F}_s$ learns global patterns of utterances commonly observable across

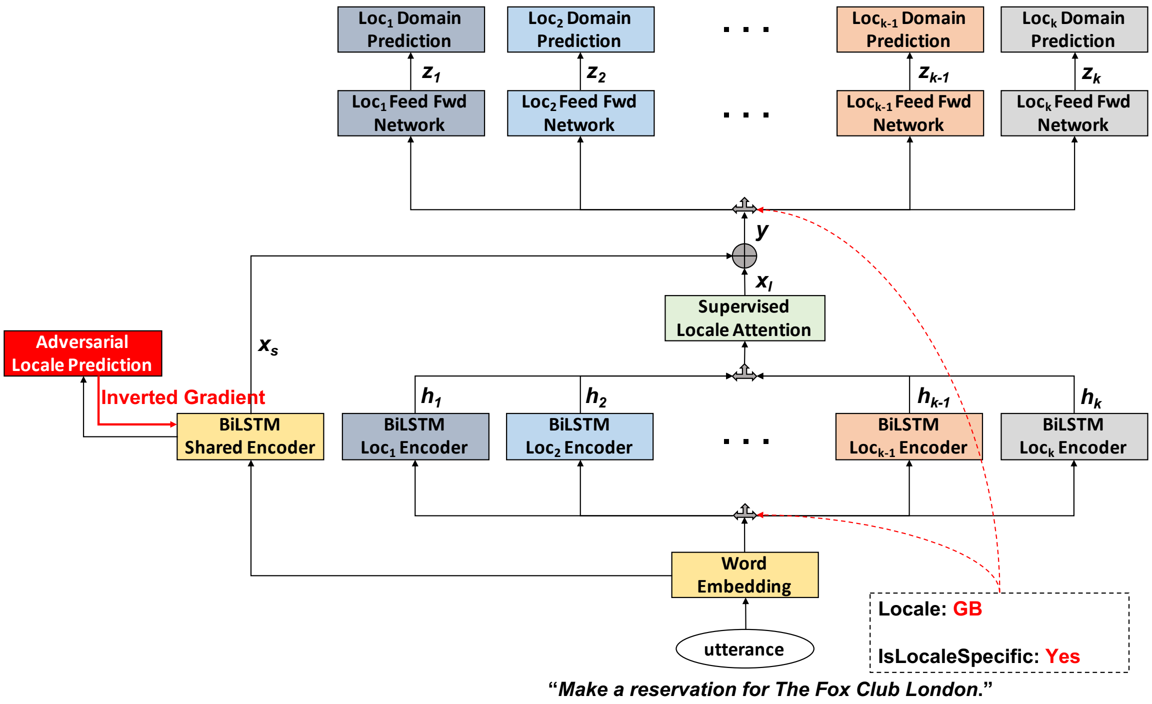

Figure 1: Model architecture of the universal model.

different locales, each of the locale-specific encoders $\mathcal{F}_{l_i}$, which corresponds to one of the locales l_i, learns local patterns of utterances that are observed specifically in the locale l_i.

3.2 Adversarial Locale Prediction Loss

Intuitively, the shared encoder $\mathcal{F}_s$ is expected to be able to better capture common utterance patterns over all locales rather than to learn patterns that are seen in only some particular locales. Thus, $\mathcal{F}_s$ can be further tuned to be locale-invariant by adding a locale prediction layer with negative gradient flow, as similarly proposed in (Kim et al., 2017c; Ganin et al., 2016; Liu et al., 2017). Let $\mathbf{x_s}$ denote an encoded vector for an input utterance produced by the shared encoder $\mathcal{F}_s$. $\mathbf{x}_s$ is then fed into a single-layer neural network to make a prediction for its corresponding locale l_i. Formally,

$$\mathbf{z}_{adv} = softmax(\mathbf{W}_{adv} \cdot \mathbf{x}_s + \mathbf{b}_{adv}) \qquad (1)$$

where $\mathbf{W}_{adv}$ and $\mathbf{b}_{adv}$ are a weight matrix and a bias term for the locale prediction layer of the feed-forward network. Since we aim to make the shared encoder $\mathcal{F}_s$ to be locale-invariant, the adversarial locale prediction loss is given by the *positive* log-likelihood:

$$\mathcal{L}_{adv} = \sum_{i=1}^{k} t_i \log[\mathbf{z}_{adv}]^i \qquad (2)$$

where t_i is a binary indicator if locale l_i is the correct prediction or not.

3.3 Supervised Locale Attention

In order to allow the locale-specific encoders to share knowledge about common domains across locales, we give a chance to learn an input utterance to any locale-specific encoders $\mathcal{F}_{l_i}$ as long as its associated domain is in D_i, except the case of locale-specific domains (i.e., `OpenTable`). Suppose $S_{d_{ij}} = \{l_w | d_{ij} \in D_w, \forall w = 1, 2, \cdots, k\}$ if d_{ij} is not locale-specific, otherwise $S_{d_{ij}} = \{l_i\}$. That is, depending on which locales a given domain is available in and whether or not it is locale-specific, its utterance needs to be selectively routed to locale-specific encoders $\mathcal{F}_{l_i}$ where $l_i \in S_{d_{ij}}$. However, we do not know a ground-truth domain associated with an input utterance during inference and it means that there is no way to do such selective routing unfortunately. Instead, we can use supervised attention mechanism to approximate the locales in which a domain is available. During training, we have each of the locale-specific encoder outputs attend each other and provide them with information about which locales should be highly attended, as explained in the following.

Let $\mathbf{H} = [\mathbf{h}_{l_1}, \mathbf{h}_{l_2}, \dots, \mathbf{h}_{l_k}] \in \mathbb{R}^{d_h \times k}$ denote a matrix of encoded vectors generated by $\mathcal{F}_{l_i}$ for $\forall i = 1, 2, \dots, k$. Then, the attention weights are obtained as follows,

$$\mathbf{a} = logistic(\mathbf{w} \cdot tanh(\mathbf{V} \cdot \mathbf{H})) \tag{3}$$

where $\mathbf{w} \in \mathbb{R}^{d_a}$ and $\mathbf{V} \in \mathbb{R}^{d_a \times d_h}$ are learnable weight parameters, and d_a is a hyperparameter we can set arbitraily. The resulted vector $\mathbf{a}$ contains attention weights in the range between 0 and 1 over the encoded vectors $\mathbf{h_{l_1}}, \dots, \mathbf{h_{l_k}}$. Then a locale-aware encoded vector $\mathbf{x_l}$ can be achieved by taking a weighted linear combination of $\mathbf{h_{l_1}}, \dots, \mathbf{h_{l_k}}$:

$$\mathbf{x}_l = \mathbf{a} \cdot \mathbf{H}^\top \tag{4}$$

The final vector representation $\mathbf{y} \in \mathbb{R}^{2 \cdot d_h}$ for the input utterance is the concatenation of two encoded vectors $\mathbf{x}_s$ and $\mathbf{x}_l$ that are produced from $\mathcal{F}_s$ and $\mathcal{F}_l$, respectively. Note we have to make sure that the proper encoders that correspond to $S_{d_{ij}}$ always get high attention weights. Thus, instead of just letting $\mathbf{V}$ and $\mathbf{w}$ be optimized during training the model, we can optimize them in a supervised way. That is, in training time, the model is aware of locales where a ground-truth domain is available. In other words, we can reward or penalize the attention weights depending on whether or not their corresponding locales have the domain of an input utterance. Therefore, the loss function for the attention weights is defined as,

$$\mathcal{L}_{loc} = -\big(\sum_{l \in S_{d_{ij}}} \log(a_l) + \sum_{l' \notin S_{d_{ij}}} \log(1 - a_{l'})\big) \tag{5}$$

3.4 Domain Classification

Once we obtain an encoded vector $\mathbf{y}$ that represents an input utterance, we feed it into prediction layers, consisting of feed forward networks, to make predictions. Since the availability of domains depends on locales, the prediction layers use the locale information associated with the utterance to route the encoded vector to only a subset of prediction layers in which the domain of the utterance is available. Then, the output vector produced by the prediction layer specifically for the locale l_i is

$$\mathbf{z}_i = \mathbf{W}_i^2 \cdot \sigma(\mathbf{W}_i^1 \cdot \mathbf{y} + \mathbf{b}_i^1) + \mathbf{b}_i^2 \tag{6}$$

where $\mathbf{W}_i$ and $\mathbf{b}_i$ are the weight and bias parameters used by the l_i specific prediction layer, and σ is an activation function for non-linearity. Since our model is structured with a multi-task learning framework to learn a joint representation across locales, we calculate $\mathbf{z}_i$ for all $l_i \in S_{d_{ij}}$ and then the predictions are made independently. Then the prediction loss is

$$\mathcal{L}_{pos} = -\log p(d_{ij}|z_i) \tag{7}$$

$$\mathcal{L}_{neg} = -\sum_{\substack{\hat{d_{ij}} \in D_i \\ \hat{d_{ij}} \neq d_{ij}}} \log p(\hat{d_{ij}}|z_i) \tag{8}$$

$$\mathcal{L}_{pred} = \frac{1}{|S_{d_{ij}}|} \sum_{l_i \in S_{d_{ij}}} (\mathcal{L}_{pos} + \mathcal{L}_{neg}) \tag{9}$$

Note that the prediction loss must be normalized by the number of locales in S_{dij} because the size of the set changes depending on how many locales has the domain associated with an input utterance and thus the number of the final prediction layer Then, the final objective function looks as follows,

$$\underset{\theta_{\mathcal{F}_s}, \theta_{\mathcal{F}_l}, \mathbf{V}, \mathbf{w}, \mathbf{W}, \mathbf{b}}{\arg\min} \quad \mathcal{L}_{adv} + \mathcal{L}_{loc} + \mathcal{L}_{pred} \tag{10}$$

where $\theta_{\mathcal{F}_s}$ and $\theta_{\mathcal{F}_l}$ are the LSTM weight parameters in the shared encoder and the locale-specific encoders, respectively.

4 Experiments

4.1 Dataset

We use a subset of the Amazon Alexa dataset that consists of utterances spoken to Alexa by real customers over four different English locales including US (United States), GB (United Kingdom), CA (Canda), IN (India). Each of the utterances is labeled with a ground-truth domain. The main objective of this experiment should be to show the effectiveness of various approaches on domain classification task under the situation where there exist multiple locales that have imbalanced data and disparate domain sets. Thus, we consider the following two aspects: 1) how differently various domain classification approaches behave depending

Locale	Train	Validation	Test	No. domains
US	173,258	24,653	122,931	177
GB	85,539	10,378	53,226	240
CA	7,113	887	4,487	51
IN	4,821	637	2,990	41

Table 1: Data statistics

Locale	Overall	Locale-specific	Locale-independent	Single-locale	Small
US	177	15	162	0	35
GB	240	16	224	82	100
CA	51	3	48	6	33
IN	41	4	37	12	20

Table 2: Test set breakdown

	US	GB	CA	IN
US	177	155	44	26
GB		240	27	23
CA			51	10
IN				41

Table 3: Domain overlaps between locales

on domains and 2) how well they can overcome the challenging issues discussed in Section 1. To this end, we categorize all domains in the dataset into four different groups.

- **Locale-specific** A set of domains which are defined with different intents/slots across locales.
- **Locale-independent** A set of domains which have exactly the same intent/slot lists across locales.
- **Single-locale** A set of domains which are available in only a single locale.
- **Small** A set of domains that lack data in a locale but have sufficient data in other locales.

Table 1 shows its brief statistics per locale, Table 2 presents the number of domains for each of four different domain categories, and Table 3 shows how many domains are overlapping between locales.

4.2 Competing Models

We compare the performances of the following five models.

- **single** A standard BiLSTM based encoder trained with only data in a particular locale.
- **union** An extension of 'single' trained with US data additionally.
- **constrained** A BiLSTM encoder trained with all locales data. It uses the locale information associated with the utterance to route the encoded utterance to only a subset of domains available in the constrained output space for the locale to make prediction (Kim et al., 2016b,a).
- **universal** This is our main contribution model described throughout the paper.
- **universal + adv** An extension of 'universal' incorporating the adversarial locale prediction loss as discussed in Section 3.2.

4.3 Domain Classification

To demonstrate the effectiveness of our model architecture especially on domains with insufficient data and/or locale-dependency, we report the classification performances of all competing models on several subsets of the dataset (four different groups presented in Section 4.1) as well as the entire data. We use classification accuracy as our main evaluation metric. The experimental results in Table 4 clearly show two major points: 1) our proposed universal model outperforms all other baselines over all locales and all domain sets, and 2) the baseline models achieve very poor accuracy especially when leveraging available data in other locales is of critical importance or when there needs to selectively share knowledge depending on the locale-specificity of a domain. If a model that shares knowledge across locales does not handle locale-specific domains carefully, its performance would deteriorate due to confusion on locale-specific patterns. The 'constrained' model uses a shared encoder and allows locales to shares its prediction layer, but it does not determine whether or not to share knowledge for each domain. As a result, its classification accuracy is only 44% for locale-specific domains and 25% for single-locale domains in the IN dataset with lack of data. Also, 'single' and 'union' models do not have any chance to learn a joint representation while sharing knowledge and thus they totally fail to make predictions correctly for locale-specific, single-locale, and small domains. In contrast, our universal model is very robust to domains with insufficient data and domains with locale-specific patterns over all locales. It proves that our approach is very effective for capturing both global and local patterns by selectively sharing domain knowledge across locales. Also, the adversarial locale prediction is only helpful for

Locale	Model	Overall	Locale-specific	Locale-independent	Single-locale	Small
US	single	70.21	54.39	69.90	–	8.18
	union	70.21	54.39	69.90	–	8.18
	constrained	74.25	76.08	74.02	–	38.30
	universal	**82.64**	88.20	**81.92**	–	**61.79**
	universal + adv	11.13	**97.51**	0.00	–	5.38
GB	single	56.02	62.81	55.09	37.81	0.00
	union	66.61	78.74	64.96	48.19	36.54
	constrained	67.82	76.83	66.60	50.51	38.04
	universal	80.06	**88.37**	78.93	**83.60**	57.96
	universal + adv	**80.52**	85.88	**79.79**	82.22	**59.52**
CA	single	43.43	3.57	43.68	0.00	0.24
	union	61.04	10.71	61.35	0.65	30.78
	constrained	76.46	67.85	76.51	39.17	55.66
	universal	**94.00**	**75.00**	**94.12**	97.74	**77.09**
	universal + adv	35.21	71.42	34.98	**98.87**	36.69
IN	single	56.25	0.00	60.46	0.00	0.00
	union	45.93	0.00	49.38	0.00	17.96
	constrained	62.64	44.71	63.98	25.94	58.64
	universal	**88.09**	**87.01**	**88.17**	80.00	**68.47**
	universal + adv	22.30	**87.01**	17.46	**82.97**	10.50

Table 4: Domain classification accuracy over different domain categories and different locales.

locale-specific and single-locale domains. That is probably because the effect of adversarial loss paradoxically makes the model rely on only the locale-specific encoders which are well-optimized for locale-specific/single-locale domains. There needs deep analysis about why it does not affect the GB locale, and we leave it as future works.

4.4 Implementation Details

All the models were optimized using a minibatch size of 64 and trained for 20 epochs by the Adam optimizer (Kingma and Ba, 2014) with initial parameter values $\eta = 1 \times 10^{-3}, \beta_1 = 0.9, \beta_2 = 0.999$. We picked the weight parameter values that achieved the best classification accuracy on the validation set to report the test set accuracy presented in Table 4. We used pre-trained word embeddings with 100 dimensionality, generated by GloVe (Pennington et al., 2014). The dimensionality of each hidden output of LSTMs is 100 for both the shared encoder $\mathcal{F}_s$ and the locale-specific encoder $\mathcal{F}_{l_i}$, and the hidden outputs of both forward LSTM and backward LSTM are concatenated, thereby the output of each BLSTM for each time step is 200. The inputs and the outputs of the BLSTMs are regularized with dropout rate 0.5 (Pham et al., 2014).

5 Conclusion

In this paper, we propose a multi-task learning based locale-agnostic universal model for domain classification task that dynamically chooses subsets of locale-specific components depending on input data. It leverages available data across locales sharing the same language to reduce the cost of scaling the domain classification model to a larger number of locales and maximize model performance even for new locales without sufficient data. The experimental results show that the universal model effectively exploits both global and local patterns and allows locales selectively share knowledge with each other. Especially, its classification performance is notable on immature locales/domains with insufficient data and locale-specific domains.

For future work, we consider adopting the proposed model architecture to multi-lingual scenario as well. The proposed model architecture is limited to supporting multiple locales using the same language only (e.g.., English in our experiments). However, voice-driven virtual assistant systems are becoming more and more popular around the world while expanding to non-English locales such as France, Italy, Spain and so on, and there could be a lot of domains built with multiple supported languages. It will definitely make the rapid scaling of a domain classification model to a large number of locales much more challenging in the future. We plan to address several issues, including but not limited to: 1) how can we capture and share knowledge of common patterns of utterances belonging to the same domain but written in different languages across different locales? 2) how can we prevent a locale from interfering with other locales using different language for learning linguistic context of utterances?

References

Isabelle Augenstein, Sebastian Ruder, and Anders Søgaard. 2018. Multi-task learning of pairwise sequence classification tasks over disparate label spaces. In *In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017a. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2832–2838.

Young-Bum Kim, Sungjin Lee, and Karl Stratos. 2017b. Onenet: Joint domain, intent, slot prediction for spoken language understanding. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pages 547–553. IEEE.

Young-Bum Kim, Alexandre Rochette, and Ruhi Sarikaya. 2016a. Natural language model reusability for scaling to different domains. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2071–2076.

Young-Bum Kim, Karl Stratos, and Dongchan Kim. 2017c. Adversarial adaptation of synthetic or stale data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1297–1307.

Young-Bum Kim, Karl Stratos, and Ruhi Sarikaya. 2016b. Domainless adaptation by constrained decoding on a schema lattice. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2051–2060.

Young-Bum Kim, Karl Stratos, and Ruhi Sarikaya. 2016c. Frustratingly easy neural domain adaptation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 387–396.

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

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

Vu Pham, Théodore Bluche, Christopher Kermorvant, and Jérôme Louradour. 2014. Dropout improves recurrent neural networks for handwriting recognition. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 285–290. IEEE.

Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Ruhi Sarikaya. 2017. The technology behind personal digital assistants: an overview of the system architecture and key components. *IEEE Signal Processing Magazine*, 34(1):67–81.

Ruhi Sarikaya, Paul A Crook, Alex Marin, Minwoo Jeong, Jean-Philippe Robichaud, Asli Celikyilmaz, Young-Bum Kim, Alexandre Rochette, Omar Zia Khan, Xiaohu Liu, et al. 2016. An overview of end-to-end language understanding and dialog management for personal digital assistants. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 391–397. IEEE.