ADVIN: AUTOMATICALLY DISCOVERING NOVEL DOMAINS AND INTENTS FROM USER TEXT UTTERANCES

Nikhita Vedula, Rahul Gupta, Aman Alok, Mukund Sridhar, Shankar Ananthakrishnan

Amazon.com, Inc. {veduln, gupra, alokaman, harakere,sanantha}@amazon.com

ABSTRACT

Recognizing the intents and domains of users' spoken and written language is a key component of Natural Language Understanding (NLU) systems. Real applications however encounter dynamic, rapidly evolving environments with newly emerging intents and domains, for which no labeled data or prior information is available. For such a setting, we propose a novel framework, *ADVIN*, to automatically discover novel domains and intents from large volumes of unlabeled text. We first employ an open classification model to discriminate all utterances potentially consisting of a novel intent. Next, we train a deep learning model with a pairwise margin loss function and knowledge transfer, to discover multiple *latent* intent categories in an unsupervised manner. We finally form a hierarchical intent-domain taxonomy by linking mutually related novel intents into novel domains. ADVIN significantly outperforms strong baselines on four benchmark datasets, and data from a real-world voice agent.

Index Terms— Intent Detection, Domain Detection, Language Understanding

1. INTRODUCTION

Comprehending the *intent* and/or *domain* (groups of mutually related intents) of users' language utterances is a key task in everyday gadgets like mobile phones or smart speaker devices. Various techniques have been proposed in the literature for this [1–11]. Most methods are supervised or semi-supervised, i.e. they capitalize on sufficient labeled data, can only handle a fixed number of intents and domains *seen* during model training, and generalize poorly to new intents or domains *unseen* during model development. Zero shot techniques [6, 12] recognize new intents for which no labeled training data is available. However, they require some (often unfeasible) additional information like the number of new intent types, and some prior knowledge about the new intents to be discovered. Efforts have been made to break the closed-world assumption in the NLU literature [9, 10, 13, 14] via *open world learning* or *open classification*, that identifies instances with labels unseen during training. However, the task of discovering the actual latent categories within the instances identified as possessing unseen labels is relatively under explored [10, 15]. In this work, we attempt to bridge the gap between the two challenging yet realistic tasks of (i) discriminating utterances belonging to new intents/domains from utterances belonging to already familiar ones, and (ii) organizing the newly discovered intents/domains into a taxonomy. Though we address the problem of novel user intent and domain discovery, our technique can easily be generalized to any open classification setting. We propose a novel, three-stage framework called *ADVIN* (Automated Discovery of noVel domaIns and iNtents). It automatically discovers user intents and domains in massive, unlabeled text corpora, *without* any prior knowledge about the intents or domains that the text may comprise of. Our method first leverages the pre-trained multi-layer transformer network, BERT [16], to determine if an utterance is likely to contain a novel intent or not. ADVIN next uses unsupervised knowledge transfer to discover the latent intent categories in the earlier identified utterances. Finally, ADVIN hierarchically links semantically related groups of discovered intents to form new domains. Thus, our main contributions are that ADVIN (i) is completely unsupervised with respect to the number and names of novel intents and domains it detects; (ii) jointly detects novel domains and novel intents to form an intent-domain taxonomy; and (iii) defines a new loss function to learn pairwise distances between utterances.

2. OUR MULTI-STAGE FRAMEWORK ADVIN

Consider a corpus of utterances $\mathcal{D}_{\mathcal{T}}$ labeled with S seen intents and S_D seen domains; and a corpus of unlabeled utterances $\mathcal{D}_{\mathcal{C}}$ consisting of U novel intents and U_D novel domains. $S \cap U = \varnothing$ and $S_D \cap U_D = \varnothing$.

Stage I: Detecting Instances with Novel Intents. We construct a two-step system to detect the presence of *novel* intents U in input instances, *without* any labeled training examples for these novel intents. Prior open classification literature has modeled the problem of finding *novel* or *unseen* classes as an $(S+1)$-class classification problem, with S *seen* classes and an additional *unseen* class [13, 15, 17]. We leverage vast amounts of linguistic and contextual knowledge from the pre-trained language model BERT [16], to learn an $(S+1)$-class classifier (Figure 1), that classifies utterances as containing a novel intent or not. We fine tune the parameters θ in different BERT layers with different learning rates

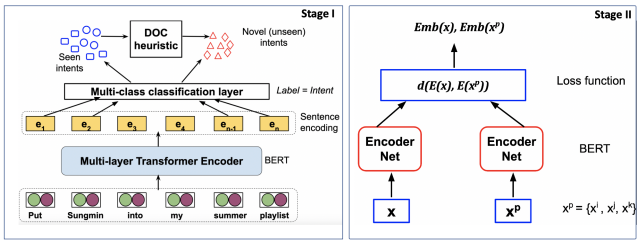

Fig. 1: Overview of both stages I (Detecting instances containing novel intents) and II (Discovering novel intent categories) of our proposed approach ADVIN

as per [18]. Labeled training data $\mathcal{D}_\mathcal{T}$ is available for the S seen intents. We use *out-of-domain* (OOD) intent detection datasets distinct from both $\mathcal{D}_\mathcal{T}$ and $\mathcal{D}_\mathcal{C}$, as training data for the $(S+1)$-th 'novel' intent class. This OOD data comes from out-of-domain, intent-labeled publicly available datasets (e.g. SNIPS [19], ATIS [20]) that do not require extra human annotation effort. Note that while training the $(S+1)$-class classifier, ADVIN only requires the information that the OOD intents do not overlap with those in $\mathcal{D}_\mathcal{T}$ (by definition), and *not* the actual intent labels of the OOD data. However, if the OOD data (e.g. SNIPS) is annotated with m intent classes, ADVIN can use this information by fragmenting the $(S+1)^{\text{th}}$ class into m classes, forming an $(S+m)$-class classifier. Using m classes may provide a better representation of the texts containing unseen intents (see Table 2). An utterance classified into any of the m classes is flagged as having a novel intent. To learn an effective distribution of the data belonging to the seen and novel intents, we adopt an additional step inspired by the Deep Open Classification algorithm (DOC) [13]. After Step I classifies the input as having a seen or novel intent, ADVIN learns statistical confidence thresholds for each seen intent s_i. It thus captures instances that have been classified to one the S seen intents with a low confidence. If the class-specific prediction probabilities for an utterance are less than the thresholds learned for each seen intent s_i, that utterance is also classified as having a novel intent (Figure 1).

Stage II: Discovering the Latent Novel Intent Categories from the Unlabeled Instances. We use complete-linkage, agglomerative hierarchical clustering [21] to group together related utterances in $\mathcal{D}_\mathcal{X}$, and find the potential novel intents U. We assume that utterances in the training set $\mathcal{D}_\mathcal{T}$ of seen intents, as well as those in the unlabeled corpus $\mathcal{D}_\mathcal{X}$ of newly emerging unseen intents come from similar distributions. First, we hierarchically cluster the labeled training data utterances $\mathcal{D}_\mathcal{T}$, using the seen intents as ground truth cluster labels. We obtain a distance value δ by maximizing the F1 score of their clustering arrangement. In an ideal scenario, every obtained cluster L_i represents a single seen intent. Our final step is then to *transfer* this distance threshold δ learnt from the *seen* intent utterances, to hierarchically cluster utterances in $\mathcal{D}_\mathcal{X}$ containing *novel* or unseen intents. This distance threshold δ is defined as $\max_{x \in L_i, x^p \in L_j} f(Emb(\boldsymbol{x}), Emb(\boldsymbol{x^p}))$ according to complete-linkage hierarchical clustering. Here, function $f(.)$ quantifies the distance between the embeddings $Emb(.)$ of an utterance pair $(\boldsymbol{x}, \boldsymbol{x^p})$ belonging to clusters L_i and L_j respectively.

We train a neural network model to obtain utterance embeddings $Emb(.)$ for both the labeled corpus of seen intents ($\mathcal{D}_\mathcal{T}$) and the unlabeled utterances detected to contain novel intents ($\mathcal{D}_\mathcal{X}$). We utilize the knowledge from both the seen intents S as well as seen domains S_D to train this model (see Figure 1). We create a training dataset from $\mathcal{D}_\mathcal{T}$ comprising pairs of utterances (x, x^p). Both x and x^p contain a seen intent. For each x, there are three possible choices for its paired utterance x^p: (i) x^i containing the same domain and intent as x, (ii) x^j containing the same domain but different intent than x, and (iii) x^k containing a different domain and intent than x. Thus, $x^p \in \{x^i, x^j, x^k\}$. These utterance pairs (x, x^p) are fed as input to a BERT transformer block, termed `EncoderNet`. We use the same learning rate decay strategy of fine-tuning the BERT layers as Stage I. Representations $E(\boldsymbol{x})$ and $E(\boldsymbol{x^p})$ are learned by the second last BERT layer. The next layer on top of the `EncoderNet` blocks uses a distance function d to compute pairwise representation distances, subject to the following bi-directional constraints: (i) the distance $d(E(x), E(x^i))$ between the representations of x and x^i should be less than $d(E(x), E(x^j))$; (ii) $d(E(x), E(x^i))$ should be less than $d(E(x), E(x^k))$; and (iii) the distance $d(E(x), E(x^j))$ between the representations of x and x^j should be less than the distance $d(E(x), E(x^k))$ between the representations of x and x^k. These constraints use linguistic and semantic relationships to ensure that utterances containing distinct domains or intents should be more distant in the learned embedding space from utterances containing the same domains or intents. We then formulate a loss function $\mathfrak{L}$ to train our model:

$$\begin{aligned}\mathfrak{L} = \frac{1}{M} \sum_{i,j,k} \{ & \max[0, m_1 + d(E(x), E(x^i)) - d(E(x), E(x^j)] \\ & + \alpha \max[0, m_2 + d(E(x), E(x^i)) - d(E(x), E(x^k)] \\ & + \beta \max[0, m_3 + d(E(x), E(x^j)) - d(E(x), E(x^k)]\}\end{aligned}$$

where m_1, m_2, m_3 are predefined margins, α and β are predefined weighting scalars and M is the total number of utterance pairs (x, x^p). We found such a loss formulation to outperform the popular contrastive loss [22] and triplet loss [23] functions (see Table 3). Next, a non-linear activation followed by a linear layer outputs embeddings $Emb(\boldsymbol{x})$ and $Emb(\boldsymbol{x^p})$ for the pair (x, x^p). Finally, pairwise distances $f(Emb(\boldsymbol{x}), Emb(\boldsymbol{x_p}))$ are computed between all utterance pairs in this embedding space, to be given as input to hierarchical clustering. Note that we do not require any intent or domain labels for the unlabeled corpus $\mathcal{D}_\mathcal{X}$ containing the unseen (novel) intents.

Stage III: Linking Mutually Related Novel Intents into Novel Domains to form a Taxonomy. We hypothesize that

using the domain labels available for the seen intents directly can better categorize the related novel intents into novel domains. We perform these steps to link intents into domains and create an intent-domain taxonomy.

(i) Assuming an ideal clustering in Stage II, each seen intent-cluster L_i will contain utterances belonging to a single seen intent $s_i \in S$. The domain label of cluster L_i would be the seen domain $\in S_D$ of intent s_i itself. However, L_i may not be completely pure, i.e. it may contain utterances with varied seen intent labels. In such cases, we assign the domain label for the seen intent-cluster L_i as the domain of the intent of the majority of the utterances in L_i.

(ii) We obtain a representation $Emb_L(L_i)$ for each seen intent cluster and each novel intent cluster, L_i, as the average of the embeddings $Emb(\boldsymbol{x})$ (Figure 1) of all utterances $x \in L_i$.

(iii) Finally, we cluster the representations $Emb_L(L_i)$ of the seen intent-clusters L_i themselves. This time we use the seen *domains* as ground truth cluster labels (instead of the seen *intents* as in Section 2). As earlier, we obtain a distance threshold δ that maximizes the F1 score for the seen domains. We transfer this threshold to perform hierarchical clustering of the novel intent-clusters. Each cluster so obtained contains groups of related novel intents, representing novel domains. Thus, ADVIN creates a taxonomy of novel intents and domains from unlabeled utterances.

3. EVALUATION

We test ADVIN on both controlled and real-world intent and domain detection datasets of SNIPS [19], ATIS [20], Facebook's task-oriented parsing (FTOP) data [24], NLU evaluation data (NLUED) [25] and Internal NLU Data from a commercial voice assistant, containing utterances of different lengths. We completely remove all utterances associated with certain random sets of intents and domains from the training and validation sets of ADVIN (Table 1). We acknowledge that this might not be completely possible in case of real applications without manual effort, since an unlabeled data corpus may consist of any number of seen and unseen intents. However, the classifier in Stage I of our approach will still be able to handle such cases by definition, since it has been trained to detect instances containing novel intents. We compare Stage I of ADVIN of detecting the presence of novel intents in input utterances, with existing methods (DOC [13], IntentCaps [6], LOF-CL [9]), as well as variants of ADVIN in Table 2.

(i) **ADVIN *(1-unseen)*/ADVIN *(1-unseen + DOC)***: variants of ADVIN using an *(S+1)*-class classifier in Stage I (Section 2), with and without the additional check per the DOC heuristic (Step II of Stage I).

(ii) **ADVIN *(m-unseen)***: using an *(S+m)*-class classifier, without the DOC heuristic (Step II) in Stage I. We utilize 'm' intent class labels (if available) for the *OOD public data* (described in Section 2. We do not require any intent labels for the unlabeled utterances $\mathcal{D}_C$.

Table 1: Evaluating ADVIN on discovering novel intents and domains removed during training.

Dataset	Sets of domains removed from training data for evaluation.	Size	Avg. text len.
SNIPS	**Set 1**: *Weather, Restaurant*; **Set 2**: *AddToPlaylist, RateBook*	13.8K	9.05
ATIS	**Set 1**: *airline, meal, airfare, day-name, distance*; **Set 2**: *flight-time, flight-no, flight, aircraft, ground-service*	5.87K	11.2
FTOP	**Set 1**: *unsupported, unsupported-event, unsupported-navigation, unintelligible*	44.78K	8.93
NLUED	**Set 1**: *Email, Cooking, Transport*; **Set 2**: *Alarm, Audio, Calendar*	11.1K	6.84
Internal NLU Data	**Set 1**: *Weather, Calendar, Todos*; **Set 2**: *Bookings, Sports, Search, Video, Media*; **Set 3**: *Recipe, Music, Shopping,Communication*	3.16M	3.72

Table 2: F1-score of various approaches for detecting if an utterance contains a novel intent (Stage I).

Approach	SNIPS		ATIS		FTOP	NLUED	Internal Dataset		
	Set1	Set2	Set1	Set2	Set1	Set1	Set1	Set2	Set3
DOC [13]	0.73	0.69	0.71	0.7	0.76	0.73	0.7	0.73	0.72
IntentCaps [6]	0.81	0.77	0.7	0.75	0.8	0.81	0.82	0.83	0.78
LOF-CL [9]	0.79	0.73	0.68	0.74	0.78	0.8	0.84	0.8	0.82
ADVIN									
(1-unseen)	0.76	0.73	0.68	0.72	0.78	0.76	0.77	0.75	0.79
(1-unseen +DOC)	0.85	0.81	0.73	0.8	0.87	0.84	0.86	0.87	0.88
(m-unseen)	0.78	0.75	0.7	0.75	0.8	0.85	0.8	0.8	0.82
(proposed)	**0.9**	**0.87**	**0.78**	**0.84**	**0.9**	**0.89**	**0.9**	**0.92**	**0.9**

We next evaluate Stages II and III of discovering the actual intent categories in user utterances identified by Stage I, and linking the newly discovered intents into domains. As per our knowledge, work in the literature only detects if utterances contain novel intents or not, and does not identify groups of latent intent categories in unlabeled text. Thus, we compare ADVIN with its own variants in Table 3:

(i) **ADVIN *(clf+hier)***: uses the representation learned by the 2nd to last BERT layer of our Stage I classification model in Figure 1, as input to hierarchical clustering.

(ii) **ADVIN *(triplet+hier)***: uses embeddings learned by a triplet network [23] as input to hierarchical clustering. The inputs to the network are utterance triplets (x, x^-, x^+). x^- and x^+ contain the same domain and intent, and different domain and intent as utterance x respectively.

(iii) **ADVIN *(ProdLDA)***: uses a neural topic modeling method ProdLDA [26] to discover novel intents, instead of clustering. ProdLDA needs the number of topics as input, so we give it the number of clusters output by "ADVIN *(as proposed)*". We choose topic modeling because it is a good, unsupervised alternative way of discovering clusters or components in data. Ground truth intent and domain labels are available for the various data Sets we used for evaluation (see Table 1). We thus evaluate the novel intents and domains discovered via standard clustering metrics: (i) Comparing the number of discovered intents and domains to the ground truth number; (ii) Normalized Mutual Information (NMI); (iii) Pu-

Table 3: Discovering the latent intent types for utterances with novel intents (Stage II). '#int.' shows the number of discovered intents, 'GT' denotes the true number of intents, and 'Pur.' denotes cluster purity.

Approach	SNIPS Set 1(GT=2) #int.,NMI,Pur.,F1	SNIPS Set 2(GT=3) #int., NMI, Pur., F1	FTOP Set 1(GT=4) #int.,NMI,Pur.F1)
Clf+hier	**3**, 0.78, 0.9, 0.76	**3**, 0.7,0.8, 0.69	24, 0.4, 0.56, 0.38
Triplet+hier	4, 0.71, 0.81, 0.69	5, 0.65, 0.76, 0.66	48, 0.36, 0.51, 0.35
ProdLDA	NA, 0.71, 0.84, 0.72	NA, 0.66, 0.79, 0.68	NA, 0.42, 0.53, 0.38
Proposed	**3**, **0.8**, **0.92**, **0.78**	**3**, **0.72**, **0.83**, **0.71**	**19**, **0.46**, **0.61**, **0.51**

Table 4: Evaluating ADVIN at predicting if a pair of utterances contain the same novel intent or not. We show the F1-score computed via two sources of ground truth (GT) intent labels, separated by a '/': (i) a user study (US) and (ii) original dataset-provided intent annotations.

Approach	FTOP: F1 US-GT / F1 Dataset-GT	Internal: F1 US-GT / F1 Dataset-GT
ADVIN *(clf+hier)*	0.66 / 0.6	0.6 / 0.68
ADVIN*(triplet+hier)*	0.6 / 0.57	0.55 / 0.61
ADVIN *(proposed)*	0.7 / 0.61	0.61 / 0.71

rity: the extent to which a cluster contains utterances having the same intent or domain label; (iv) F1 score.

Table 2 shows the F1-score of various approaches, on classifying an intent as novel or not on different datasets with different dataset configurations. We observe that our proposed approach in the last row, using an $(S + m)$-class classifier and the DOC heuristic, significantly outperforms all baselines on all datasets by at least 6% F1 score points. ADVIN also outperforms the zero shot IntentCaps model (that uses relevant information available beforehand about the new test intents to be discovered), *without* using any prior knowledge about the novel intents to be discovered. The last row of Table 3 shows that Stage II of ADVIN as proposed significantly outperforms all baselines. The learned intent-clusters have a purity $>$75% and F1-score $>$60% for the SNIPS, NLUED and Internal datasets. Purity decreases to 61% for FTOP, primarily because semantically diverse utterances have the same ground truth intent label of *'unsupported'* or *'unsupported-event'*. The 'NA' value for "ADVIN *(ProdLDA)*" in the column '# int.' denotes that the *ProdLDA* algorithm does not output the number of new intents discovered. It takes this as input in the form of the number of topics. We find similar empirical trends for ATIS Sets 1 and 2, NLUED Set 2 and Internal Data Sets 2, 3 and 4. ADVIN discovers 1.5-4.5 times more novel intents than present in the ground truth, as seen from the '# int.' columns in Table 3. For example, we observe that the *'unsupported'* category of FTOP has been split up by ADVIN into 7 finer-grained, semantically sensible novel intents. E.g., the utterances *"What city has the most traffic in the US"* and *"Family friendly bars near me"*, from the *'unsupported'* intent category, have been separated by ADVIN into different intent categories. We thus find that ADVIN largely learns semantically appropriate, discriminative representations for the novel intents.

We next perform manual evaluation of ADVIN by recruiting crowd workers on Amazon Mechanical Turk for the FTOP data and employees familiar with the Internal Dataset. We provide a set of random utterance pairs predicted by ADVIN as having novel intents to the annotators, and ask them to indicate whether the pair is likely to belong to the same intent category or not. Each instance is annotated by three annotators, and we followed recommended practices for crowd sourcing [27] to ensure good quality annotations. For both datasets we compute the F1-score in Table 4, by comparing the output of ADVIN with that of the human annotators, for 2500 FTOP utterance pairs (inter-annotator agreement Cohen's $\kappa = 0.78$) and 1100 Internal data utterance pairs (Cohen's $\kappa = 0.9$). We observe that ADVIN as proposed, significantly outperforms baselines by at least 5% with respect to human evaluation.

Table 5: Sample intents and domains detected by ADVIN.

FTOP data utterances containing novel intents FI1, FI2, FI3	NLUED utterances containing novel intents NI1,NI2,NI3
FI1: *Any parks near Fillmore that offer sledding; Do they have snow sleds at Ober Gatlinburg; Ice skate rink hours for Dec 9th* **FI2**: *Closest nightclub that has dancing; Find the nearest dance club that has a live band; Directions to the Die Antword concert* **FI3**: *Are there any drive-in movie theaters left in Ohio that are open in the fall; Movie theaters near me*	**NI1**: *Tell me about Mary S. in my contacts; how many numbers are saved for Ale; what is John Doe's address* **NI2**: *i need an email chain to my mother i'm planning a trip to see her ask her how the weather will be so i know how to pack; what is the subject of the email that just arrived; check for this mail in my contact if not then add it*
Novel domain FD1: novel intents FI2 and FI3	**Novel domain ND1**: novel intents NI1 and NI2

We show in Table 5 sample utterances that have been discovered by ADVIN as containing novel intents and domains. Utterances talking about similar or related topics (e.g. *sledding*, *ice skating* in the first row) are grouped into a single intent category. In the second column, utterances about various *email* aspects are collated into a novel 'email' based domain.

4. CONCLUSION

We propose a generalizable framework, ADVIN, to discover novel intents and novel domains in unlabeled text, utilizing an existing labeled corpus of non-overlapping intents, a pre-trained language model and knowledge transfer. We first identify all unlabeled utterances containing novel intents. We develop a model that maximizes inter-intent variance and minimizes intra-intent variance between labeled utterance pairs; and transfers knowledge learned from the labeled intents seen during training to the unlabeled data containing novel intents. We also form an intent-domain taxonomy by hierarchically linking the detected, related novel intents into novel domains. ADVIN significantly outperforms strong baselines on real data. In future, we aim to detect multiple intents or domains per utterance; and use additional knowledge to better model human-perceived intents and domains. A more detailed version of this work is available here [28].

5. REFERENCES

[1] Gokhan Tur, Dilek Hakkani Tür, Larry Heck, and Sarangarajan Parthasarathy, "Sentence simplification for spoken language understanding," in *ICASSP*, 2011.

[2] Minwoo Jeong and Gary Geunbae Lee, "Triangular-chain conditional random fields," *IEEE TASLP*, 2008.

[3] Joo-Kyung Kim, Gokhan Tur, Asli Celikyilmaz, Bin Cao, and Ye-Yi Wang, "Intent detection using semantically enriched word embeddings," in *IEEE SLT*, 2016.

[4] Suman Ravuri and Andreas Stoicke, "A comparative study of neural network models for lexical intent classification," in *IEEE ASRU*, 2015.

[5] Ming Sun, Aasish Pappu, Yun-Nung Chen, and Alexander I Rudnicky, "Weakly supervised user intent detection for multi-domain dialogues," in *IEEE SLT*, 2016.

[6] Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S Yu, "Zero-shot user intent detection via capsule neural networks," *arXiv preprint arXiv:1809.00385*, 2018.

[7] Prashanth Gurunath Shivakumar, Mu Yang, and Panayiotis Georgiou, "Spoken language intent detection using confusion2vec," *arXiv:1904.03576*, 2019.

[8] Giuseppe Castellucci, Valentina Bellomaria, Andrea Favalli, and Raniero Romagnoli, "Multi-lingual intent detection and slot filling in a joint bert-based model," *arXiv:1907.02884*, 2019.

[9] Ting-En Lin and Hua Xu, "Deep unknown intent detection with margin loss," *arXiv preprint arXiv:1906.00434*, 2019.

[10] Nikhita Vedula, Nedim Lipka, Pranav Maneriker, and Srinivasan Parthasarathy, "Open intent extraction from natural language interactions," in *WWW*, 2020.

[11] Momchil Hardalov, Ivan Koychev, and Preslav Nakov, "Enriched pre-trained transformers for joint slot filling and intent detection," *arXiv:2004.14848*, 2020.

[12] Anjishnu Kumar, Pavankumar Reddy Muddireddy, Markus Dreyer, and Björn Hoffmeister, "Zero-shot learning across heterogeneous overlapping domains.," in *INTERSPEECH*, 2017.

[13] Lei Shu, Hu Xu, and Bing Liu, "Doc: Deep open classification of text documents," *arXiv preprint arXiv:1709.08716*, 2017.

[14] Joo-Kyung Kim and Young-Bum Kim, "Joint learning of domain classification and out-of-domain detection with dynamic class weighting for satisficing false acceptance rates," *arXiv:1807.00072*, 2018.

[15] Lei Shu, Hu Xu, and Bing Liu, "Unseen class discovery in open-world classification," *arXiv preprint arXiv:1801.05609*, 2018.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[17] Hu Xu, Bing Liu, Lei Shu, and P Yu, "Open-world learning and application to product classification," in *WWW*, 2019.

[18] Jeremy Howard and Sebastian Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018.

[19] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, et al., "Snips voice platform," *arXiv preprint arXiv:1805.10190*, 2018.

[20] Deborah A Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, et al., "Expanding the scope of the atis task: The atis-3 corpus," in *HLT Workshop*, 1994.

[21] K Gowda and G Krishna, "Agglomerative clustering using the concept of mutual nearest neighbourhood," *Pattern recognition*, 1978.

[22] Raia Hadsell, Sumit Chopra, and Yann LeCun, "Dimensionality reduction by learning an invariant mapping," in *CVPR*, 2006.

[23] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[24] Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis, "Semantic parsing for task oriented dialog using hierarchical representations," *arXiv preprint arXiv:1810.07942*, 2018.

[25] Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser, "Benchmarking natural language understanding services for building conversational agents," *arXiv preprint arXiv:1903.05566*, 2019.

[26] Akash Srivastava and Charles Sutton, "Autoencoding variational inference for topic models," *arXiv preprint arXiv:1703.01488*, 2017.

[27] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality data?," 2016.

[28] Nikhita Vedula, Rahul Gupta, Aman Alok, and Mukund Sridhar, "Automatic discovery of novel intents & domains from text utterances," *arXiv:2006.01208*, 2020.