MITIGATING THE IMPACT OF SPEECH RECOGNITION ERRORS ON SPOKEN QUESTION ANSWERING BY ADVERSARIAL DOMAIN ADAPTATION

Chia-Hsuan Lee, Yun-Nung Chen, Hung-Yi Lee

College of Electrical Engineering and Computer Science
National Taiwan University, Taiwan
chiahsuan.li@gmail.com, y.v.chen@ieee.org, tlkagkb93901106@gmail.com

ABSTRACT

Spoken question answering (SQA) is challenging due to complex reasoning on top of the spoken documents. The recent studies have also shown the catastrophic impact of automatic speech recognition (ASR) errors on SQA. Therefore, this work proposes to mitigate the ASR errors by aligning the mismatch between ASR hypotheses and their corresponding reference transcriptions. An adversarial model is applied to this domain adaptation task, which forces the model to learn domain-invariant features the QA model can effectively utilize in order to improve the SQA results. The experiments successfully demonstrate the effectiveness of our proposed model, and the results are better than the previous best model by 2% EM score.

Index Terms— adversarial learning, spoken question answering, SQA, domain adaptation

1. INTRODUCTION

Question answering (QA) has drawn a lot of attention in the past few years. QA tasks on images [1] have been widely studied, but most focused on understanding text documents [2]. A representative dataset in text QA is SQuAD [2], in which several end-to-end neural models have accomplished promising performance [3]. Although there is a significant progress in machine comprehension (MC) on text documents, MC on spoken content is a much less investigated field. In spoken question answering (SQA), after transcribing spoken content into text by automatic speech recognition (ASR), typical approaches use information retrieval (IR) techniques [4] to find the proper answer from the ASR hypotheses. One attempt towards QA of spoken content is TOEFL listening comprehension by machine [5]. TOEFL is an English examination that tests the knowledge and skills of academic English for English learners whose native languages are not English. Another SQA corpus is Spoken-SQuAD[6], which is automatically generated from SQuAD dataset through Google Text-to-Speech (TTS) system. Recently ODSQA, a SQA corpus recorded by real speakers, is released [7].

To mitigate the impact of speech recognition errors, using sub-word units is a popular approach for speech-related downstream tasks. It has been applied to spoken document retrieval [8] and spoken term detection [9] The prior work showed that, using phonectic sub-word units brought improvements for both Spoken-SQuAD and ODSQA [6].

Instead of considering sub-word features, this paper proposes a novel approach to mitigate the impact of ASR errors. We consider reference transcriptions and ASR hypotheses as two domains, and adapt the source domain data (reference transcriptions) to the target domain data (ASR hypotheses) by projecting these two domains in the shared common space. Therefore, it can effectively benefit the SQA model by improving the robustness to ASR errors in the SQA model.

Domain adaptation has been successfully applied on computer vision [10] and speech recognition [11]. It is also widely studied on NLP tasks such as sequence tagging and parsing [12, 13, 14]. Recently, adversarial domain adaptation has already been explored on spoken language understanding (SLU). Liu and Lane learned domain-general features to benefit from multiple dialogue datasets [15]; Zhu et al. learned to transfer the model from the transcripts side to the ASR hypotheses side [16]; Lan et al. constructed a shared space for slot tagging and language model [17]. This paper extends the capability of adversarial domain adaptation for SQA, which has not been explored yet.

2. SPOKEN QUESTION ANSWERING

In SQA, each sample is a triple, (q, d, a), where q is a *question* in either spoken or text form, d is a *multi-sentence spoken-form document*, and a is the *answer* in text from. The task of this work is extractive SQA; that means a is a word span from the reference transcription of d. An overview framework of SQA is shown in Figure 1. In this paper, we frame the source domain as reference transcriptions and the target domain as ASR hypotheses. Hence, we can collect source domain data more easily, and adapt the model to the target domain.

In this task, when the machine is given a spoken document, it needs to find the answer of a question from the spo-

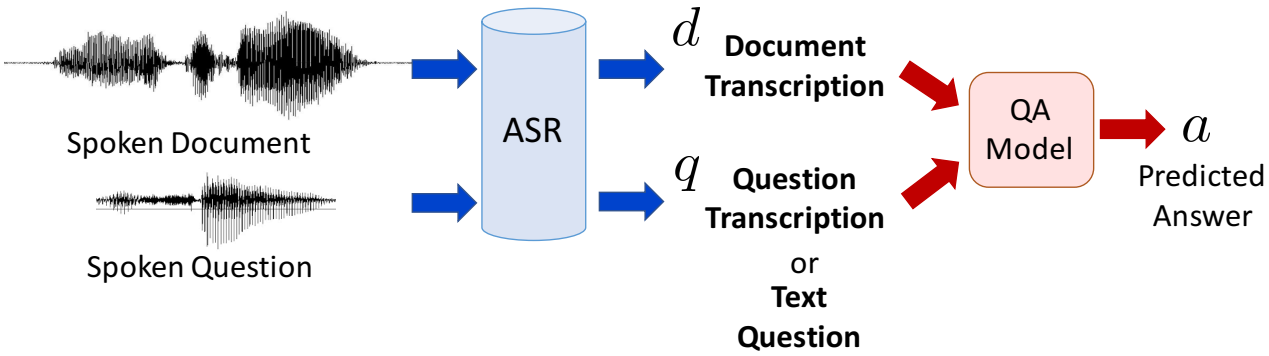

Fig. 1. Flow diagram of the SQA system.

ken document. SQA can be solved by the concatenation of an ASR module and a question answering module. Given the ASR hypotheses of a spoken document and a question, the question answering module can output a text answer.

The most intuitive way to evaluate the text answer is to directly compute the **Exact Match (EM)** and **Macro-averaged F1 scores (F1)** between the predicted text answer and the ground-truth text answer. We used the standard evaluation script from SQuAD [2] to evaluate the performance.

3. QUESTION ANSWERING MODEL

The used architecture of the QA model is briefly summarized below. Here we choose QANet [3] as the base model due to the following reasons: 1) it achieves the second best performance on SQuAD, and 2) since there are completely no recurrent networks in QANet, its training speed is 5x faster than BiDAF [18] when reaching the same performance on SQuAD.

The network architecture is illustrated in Figure 2. The left blocks and the right blocks form two QANets, each of which takes a document and a question as the input and outputs an answer. In QANet, firstly, an *embedding encoder* obtains word and character embeddings for each word in q or d and then models the temporal interactions between words and refines word vectors to contextualized word representations. All encoder blocks used in QANet are composed exclusively of depth-wise separable convolutions and self-attention. The intuition here is that convolution components can model local interactions and self-attention components focus on modeling global interactions. The *context-query attention layer* generates the question-document similarity matrix and computes the question-aware vector representations of the context words. After that, a *model encoder layer* containing seven encoder blocks captures the interactions among the context words conditioned on the question. Finally, the *output layer* predicts a start position and an end position in the document to extract the answer span from the document.

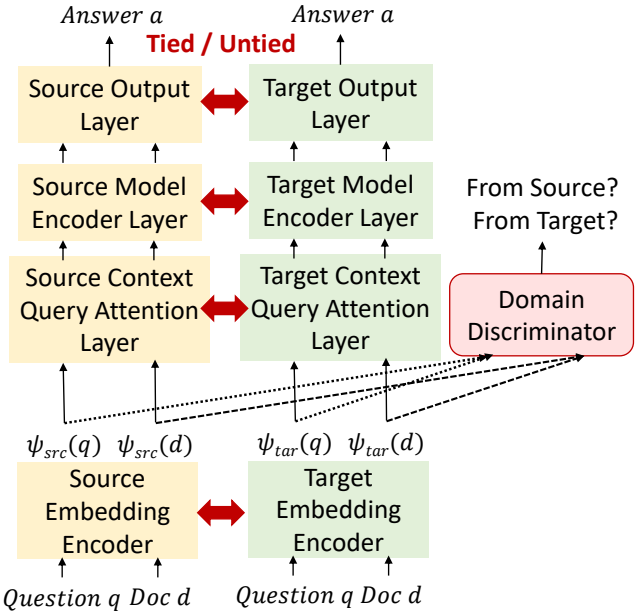

Fig. 2. The overall architecture of the proposed QA model with a domain discriminator. Each layer can be tied or untied between the source and target models.

4. DOMAIN ADAPTATION APPROACH

The main focus of this paper is to apply domain adaptation for SQA. In this approach, we have two SQA models (QANets), one trained from target domain data (ASR hypotheses) and another trained from source domain data (reference transcriptions). Because the two domains share common information, some layers in these two models can be tied in order to model the shared features. Hence, we can choose whether each layer in the QA model should be shared. Tying the weights between the source layer and the target layer in order to learn a symmetric mapping is to project both source and target domain data to a shared common space. Different combinations will be investigated in our experiments.

More specifically, we incorporate a domain discriminator into the SQA model shown in Figure 2, which can enforce the embedding encoder to project the sentences from both source and target domains into a shared common space and consequentially to be ASR-error robust. Although the embedding encoder for both domains may *implicitly* learn some common latent representations, adversarial learning can provide a more *direct* training signal for aligning the output distribution of the embedding encoder from both domains. The embedding encoder takes in a sequence of word vectors and generates a sequence of hidden vectors with the same length. We use $\Psi_{\text{tar}}(q)$ and $\Psi_{\text{tar}}(d)$ ($\Psi_{\text{src}}(q)$ and $\Psi_{\text{src}}(d)$) to represent the hidden vector sequence given the question q and the document d in the target (source) domain respectively.

The domain discriminator D focuses on identifying the domain of the vector sequence is from given Ψ_{tar} or Ψ_{src}, where the objective is to minimize L_{dis}.

$$L_{\text{dis}} = E_{(q,d,a)\sim\text{tar}}\left[\log D(\Psi_{\text{tar}}(q)) + \log D(\Psi_{\text{tar}}(d))\right] \quad (1)$$
$$+E_{(q,d,a)\sim\text{src}}\left[\log(1 - D(\Psi_{\text{src}}(q)) + \log(1 - D(\Psi_{\text{src}}(d))\right].$$

Given a training example from the target domain ($(q, d, a) \sim$ tar), D learns to assign a lower score to q and d in that example, that is, to minimize $D(\Psi_{\text{tar}}(q))$ and $D(\Psi_{\text{tar}}(d))$. On the other hand, given a training example from the source domain ($(q, d, a) \sim$ src), D learns to assign a larger value to q and d.

Furthermore, we update the parameters of the embedding encoders to maximize the domain classification loss L_{dis}, which works adversarially towards the domain discriminator. We thus expect the model to learn features and structures that can generalize across domains when the outputs of Ψ_{src} are indistinguishable from the outputs of Ψ_{tar}. The loss function for embedding encoder, L_{enc}, is formulated as

$$L_{\text{enc}} = L_{\text{qa}} - \lambda_G L_{\text{dis}}, \quad (2)$$

where λ_G is a hyperparameter. The two embedding encoders in the QA model are learned to maximize L_{dis} while minimizing the loss for QA, L_{qa}. Because the parameters of other layers in QA model are independent to the loss of the domain discriminator, the loss function of other layers, L_{other}, is equivalent to L_{qa}, that is, $L_{\text{other}} = L_{\text{qa}}$.

Although the discriminator is applied to the output of embedding encoder in Figure 2, it can be also applied to other layers.[1] Considering that almost all QA model contains such embedding encoders, the proposed approach is expected to generalize to other QA models in addition to QANet.

5. EXPERIMENTS

5.1. Corpus

Spoken-SQuAD is chosen as the target domain data for training and testing. Spoken-SQuAD [6] is an automatically generated corpus in which the document is in spoken form and the question is in text form. The reference transcriptions are from SQuAD [2]. There are 37,111 and 5,351 question answer pairs in the training and testing sets respectively, and the word error rate (WER) of both sets is around 22.7%.

The original SQuAD, Text-SQuAD, is chosen as the source domain data, where only question answering pairs appearing in Spoken-SQuAD are utilized. In our task setting, during training we train the proposed QA model on both Text-SQuAD and Spoken-SQuAD training sets. While in the testing stage, we evaluate the performance on Spoken-SQuAD testing set.

Table 1. Illustration of domain mismatch, where the models are trained on the source domain (Text-SQuAD; T-SQuAD) or the target domain (Spoken-SQuAD; S-SQuAD) and then evaluated on both source and target domains.

Model Training		T-SQuAD EM	T-SQuAD F1	S-SQuAD EM	S-SQuAD F1
T-SQuAD	(a)	61.31	72.66	42.27	55.61
S-SQuAD	(b)	45.52	57.39	48.93	61.20
Finetune	(c)	54.83	66.45	49.60	61.85

5.2. Experiment Setup

We utilize `fasttext` [19] to generate the embeddings of all words from both Text-SQuAD and Spoken-SQuAD. We adopt the phoneme sequence embeddings to replace the original character sequence embeddings using the method proposed by Li et al. [6]. The source domain model and the target domain model share the same set of word embedding matrix to improve the alignment between these two domains.

W-GAN is adopted for our domain discriminator [20], which stacks 5 residual blocks of 1D convolutional layers with 96 filters and filter size 5 followed by one linear layer to convert each input vector sequence into one scalar value.

All models used in the experiments are trained with batch size 20, using `adam` with learning rate $1e-3$ and the early stop strategy. The dimension of the hidden state is set to 96 for all layers, and the number of self-attention heads is set to 2. The setup is slightly different but better than the setting suggested by the original QAnet.

5.3. Results

5.3.1. Domain Mismatch

First, we highlight the domain mismatch phenomenon in our experiments shown in Table 1. Row (a) is when QANet is trained on Text-SQuAD, row (b) is when QANet is trained on Spoken-SQuAD, and row (c) is when QANet is trained on Text-SQuAD and then finetuned on Spoken-SQuAD. The

[1] In the experiments, we found that applying the domain discriminator to embedding encoders yielded the best performance.

Table 2. The EM/F1 scores of proposed adversarial domain adaptation approaches over Spoken-SQuAD.

Model		EM	F1
Baseline			
S-SQuAD	(a)	48.93	61.20
Finetune	(b)	49.60	61.85
Li et al. [6]	(c)	49.07	61.16
Adverarial			
Lan et al. [17]	(d)	49.13	61.80
Completely Shared	(e)	49.57	61.48
(e) + GAN on Embedding	(f)	**51.10**	**63.11**
(e) + GAN on Attention	(g)	48.30	61.11

columns show the evaluation on the testing sets of Text-SQuAD and Spoken-SQuAD.

It is clear that the performance drops a lot when the training and testing data mismatch, indicating that model training on ASR hypotheses can not generalize well on reference transcriptions. The performance gap is nearly 20% F1 score (72% to 55%). The row (c) shows the improved performance when testing on S-SQuAD due to the transfer learning via fine-tuning.

5.3.2. Effectiveness of Adversarial Domain Adaptation

To better demonstrate the effectiveness of the proposed model, we compare with baselines and show the results in Table 2. The baselines are: (a) trained on S-SQuAD, (b) trained on T-SQuAD and then fine-tuned on S-SQuAD, and (c) previous best model trained on S-SQuAD [6] by using Dr.QA [21]. We also compare to the approach proposed by Lan et al. [17] in the row (d). This approach is originally proposed for spoken language understanding, and we adopt the same approach on the setting here. The approach models domain-specific features from the source and target domains separately by two different embedding encoders with a shared embedding encoder for modeling domain-general features. The domain-general parameters are adversarially trained by domain discriminator.

Row (e) is the model that the weights of all layers are tied between the source domain and the target domain. Row (f) uses the same architecture as row (e) with an additional domain discriminator applied to the embedding encoder. It can be found that row (f) outperforms row (e), indicating that the proposed domain adversarial learning is helpful. Therefore, our following experiments contain domain adversarial learning. The proposed approach (row (f)) outperforms previous best model (row (c)) by 2% EM score and over 1.5% F1 score. We also show the results of applying the domain discriminator to the top of context query attention layer in row (g), which obtains poor performance. To sum it up, incorporating adversarial learning by applying the domain discriminator on top of the embedding encoder layer is effective.

Table 3. Investigation of different layer tying mechanisms, where ✓means that weights of the layer are tied between the source model and the target model. (L1: embedding encoder, L2: context query attention layer, L3: model encoder layer, L4: output layer.)

Combination	L1	L2	L3	L4	EM	F1
(a)	✓	✓	✓	✓	**51.10**	**63.11**
(b)	-	✓	✓	✓	50.25	62.41
(c)	-	-	✓	✓	49.72	61.97
(d)	-	✓	-	✓	48.83	61.80
(e)	-	✓	✓	-	51.09	62.97
(f)	✓	-	-	✓	49.01	61.40
(g)	✓	-	✓	-	49.28	61.71
(h)	✓	✓	-	-	49.61	61.72

5.3.3. Which Layer to Share?

Layer weight tying or untying within the model indicates different levels of symmetric mapping between the source and target domains. Different combinations are investigated and shown in Table 3. The row (a) in which all layers are tied is the row (e) of Table 2. The results show that untying context-query attention layer L2 (rows (c, f, g)) or model encoder layer L3 (rows (d, f, h)) lead to degenerated solutions in comparison to row (a) where all layers are tied. Untying both of them simultaneously leads to the worst performance which is even worse than the finetuning (row (g) v.s. (c) from Table 2). These results imply that sharing the *context-query attention layer* and the *model encoder layer* are important for domain adaptation on SQA. We conjecture that these two layers benefit from training on source domain data where there are no ASR errors, so the QA model learns to conduct attention or further reason well on target domain data with ASR errors.

Overall, it is not beneficial to untie any layer, because no information can be shared across different domains. Untying the embedding encoder L1 and the output layer L4 leads to the least degradation in comparison to row (a).

6. CONCLUSION

In this work, we incorporate a domain discriminator to align the mismatched domains between ASR hypotheses and reference transcriptions. The adversarial learning allows the end-to-end QA model to learn domain-invariant features and improve the robustness to ASR errors. The experiments demonstrate that the proposed model successfully achieves superior performance and outperforms the previous best model by 2% EM score and over 1.5% F1 score.

7. REFERENCES

[1] C Lawrence Zitnick, Ramakrishna Vedantam, and Devi Parikh, "Adopting abstract images for semantic scene understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 4, pp. 627–638, 2016.

[2] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

[3] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le, "Qanet: Combining local convolution with global self-attention for reading comprehension," *arXiv preprint arXiv:1804.09541*, 2018.

[4] Sz-Rung Shiang, Hung-yi Lee, and Lin-shan Lee, "Spoken question answering using tree-structured conditional random fields and two-layer random walk," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[5] Bo-Hsiang Tseng, Sheng-Syun Shen, Hung-Yi Lee, and Lin-Shan Lee, "Towards machine comprehension of spoken content: Initial toefl listening comprehension test by machine," *arXiv preprint arXiv:1608.06378*, 2016.

[6] Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee, "Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension," *arXiv preprint arXiv:1804.00320*, 2018.

[7] Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-Yi Lee, "Odsqa: Open-domain spoken question answering dataset," *arXiv preprint arXiv:1808.02280*, 2018.

[8] Kenney Ng and Victor W Zue, "Subword unit representations for spoken document retrieval," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[9] Charl van Heerden, Damianos Karakos, Karthik Narasimhan, Marelie Davel, and Richard Schwartz, "Constructing sub-word units for spoken term detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5780–5784.

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[11] Yusuke Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition.," in *INTERSPEECH*, 2016, pp. 2369–2372.

[12] Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen, "Transfer learning for sequence tagging with hierarchical recurrent networks," *arXiv preprint arXiv:1703.06345*, 2017.

[13] David McClosky, Eugene Charniak, and Mark Johnson, "Automatic domain adaptation for parsing," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 28–36.

[14] Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan, "Domain adaptation of rule-based annotators for named-entity recognition tasks," in *Proceedings of the 2010 conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2010, pp. 1002–1012.

[15] Bing Liu and Ian Lane, "Multi-domain adversarial learning for slot filling in spoken language understanding," *arXiv preprint arXiv:1711.11310*, 2017.

[16] Su Zhu, Ouyu Lan, and Kai Yu, "Robust spoken language understanding with unsupervised asr-error adaptation," 2018.

[17] Ouyu Lan, Su Zhu, and Kai Yu, "Semi-supervised training using adversarial multi-task learning for spoken language understanding," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6049–6053.

[18] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi, "Bidirectional attention flow for machine comprehension," *arXiv preprint arXiv:1611.01603*, 2016.

[19] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[20] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.

[21] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes, "Reading wikipedia to answer open-domain questions," *arXiv preprint arXiv:1704.00051*, 2017.