# DOMAIN ADAPTATION WITH AUGMENTED SPACE METHOD FOR MULTI-DOMAIN CONTACT CENTER DIALOGUE SUMMARIZATION

*Hitoshi Nishikawa, Toshiro Makino and Yoshihiro Matsuo*

NTT Cyber Space Labs, NTT Corp, Yokosuka, Kanagawa, Japan.
{ nishikawa.hitoshi, makino.toshiro, matsuo.yoshihiro } @lab.ntt.co.jp

## ABSTRACT

In this paper we propose a method to improve the quality of extractive summarization for contact center dialogues in various domains by making use of training samples whose domains are different from that of the test samples. Since preparing sufficient numbers of training samples for each domain is too expensive, we leverage references from many different domains and employ the Augmented Space Method to implement domain adaptation. As the target of summarization, we take up contact center dialogues in six domains and summarize their transcripts. Our experiment shows that the proposed method achieves better results than the usual supervised learning approach.

***Index Terms***— speech summarization, domain adaptation

## 1. INTRODUCTION

Contact center dialogue summarization is attracting much more attention [1, 2, 3]. If contact center dialogues can be summarized automatically, business enterprises can extract valuable information from the summaries and leverage the data to improve their businesses and make better decisions.

Implementing an automatic summarization system that outputs good summaries requires the manual estimation of parameters from a set of pairs of documents and their references [4]. If there is a sufficient number of training samples, the summarizer can learn what summaries are expected. That is, a summarizer can generate good summaries if it uses the parameters estimated from a sufficient number of training samples.

However, preparing a sufficient number of training examples is expensive. In addition, the properties of the desired summaries depend largely on the domain of the input documents; therefore many training samples of different domains must be made to provide ensure adequate coverage. For the example of contact center dialogue summarization, dialogues in the contact center of a bank and those in the center of an internet service provider should differ quite a bit. Therefore, training samples must be made for both domains which incurs a lot of cost.

To solve this problem, in this paper, we leverage a domain adaptation technique which uses training samples whose domain is the same as that of the input documents and those whose domains are different from that of the input documents. For example, to summarize the contact center dialogues in the bank domain, the summarizer uses training samples in that domain and those in other domains, such as internet service provider. We adopt the Augmented Space Method [5], a well-known domain adaptation method, to implement our approach.

We perform experiments to validate its efficacy. Our proposed method surpasses the well-known supervised approach; when training samples from different domains exist, our experiments show that domain adaptation yields the best results.

## 2. RELATED WORK

Some papers have presented methods for summarizing contact center dialogues [1, 2, 3]. Byrd et al. suggested the use of some heuristic rules to summarize contact center dialogues [1]. Higashinaka et al. train HMMs that detect characteristic utterances in the target domains and then uses these HMMs to summarize dialogues by labeling utterances [2, 3]. In contrast, our proposal is the only one to leverage the training samples of different domains.

Although target documents are not contact center dialogues, as the closest work to this paper, Sandu et al. leverage references of meeting speeches to summarize threads of e-mails [6]. However, their experiment showed that the conventional supervised learning, which uses only training samples whose domain matches that of the target documents, wins against domain adaptation methods. They said that this result is due to the wide difference between the properties of e-mails and meeting speeches. In contrast, we show that our proposed approach works well in the task of contact center dialogue summarization, though we also show that adapting largely different training samples is difficult as they pointed out.

## 3. SUMMARIZATION MODEL

In this paper, we denote a set of sentences to be summarized by $\mathbf{x}$ and its subset that meets the given length of summary by $\mathbf{y}$. Also, we denote an objective function by the function that maps summary $\mathbf{y}$ to a real number $\mathbf{f}_{\mathbf{x},\mathbf{w}} : \mathbf{y} \mapsto \mathbb{R}$ under the given sentences to be summarized $\mathbf{x}$ and parameter vector $\mathbf{w}$. In this setting, the summarization problem can be described as follows:

$$\hat{\mathbf{y}} = \underset{\mathbf{y} \subseteq \mathbf{x}}{\operatorname{argmax}} \mathbf{f}_{\mathbf{x},\mathbf{w}}(\mathbf{y}) \quad (1)$$

$$\text{s.t.} \;\; \operatorname{length}(\mathbf{y}) \leq L$$

where $\operatorname{length}$ is a function that returns the length of summary $\mathbf{y}$, $L$ is the maximum summary length.

Some previous work has shown the efficacy of the objective function that scores sentences and words separately for speech summarization [3, 7]. We adopt this kind of objective function and define our objective function as follows:

**Table 1**. Features for sentences.

| Features | Value |
|---|---|
| Normalized sentence position | [0,1] |
| # of tokens in sentence | Integer |
| # of words in sentence | Integer |

**Table 2**. Features for words.

| Features | Value |
|---|---|
| Surface of word | $\{0, 1\}$ |
| POS of word | $\{0, 1\}$ |
| Word frequency in input document | Integer |
| # of sentences containing word | Integer |

$$\mathbf{f}_{\mathbf{x},\mathbf{w}}(\mathbf{y}) = \sum_{x_i \in \mathbf{y}} \mathbf{u}^\top \phi(x_i) + \sum_{z_j \in \mathbf{y}} \mathbf{v}^\top \psi(z_j) \qquad (2)$$

where $x_i$ is the $i$ th sentence present in summary $\mathbf{y}$, $z_j$ is the $j$ th word present in summary $\mathbf{y}$. $\mathbf{u} \in \mathbb{R}^{d_u}$ and $\mathbf{v} \in \mathbb{R}^{d_v}$ are parameter vectors for sentences and words, respectively. $\phi : x \mapsto \mathbb{R}^{d_u}$ and $\psi : z \mapsto \mathbb{R}^{d_v}$ are feature functions for sentences and words, respectively. We show features for sentences in Table 1 and features for words in Table 2.

Two terms in Eq. 2, $\sum_{x_i \in \mathbf{y}} \phi(x_i)$ and $\sum_{z_j \in \mathbf{y}} \psi(z_j)$ can be represented together as $\mathbf{\Phi}(\mathbf{x}, \mathbf{y})$, also $\mathbf{u}$ and $\mathbf{v}$ can be merged as $\mathbf{w}^\top = \langle \mathbf{u}^\top, \mathbf{v}^\top \rangle$. Hence our objective function can be represented as $\mathbf{f}_{\mathbf{x},\mathbf{w}}(\mathbf{y}) = \mathbf{w}^\top \mathbf{\Phi}(\mathbf{x}, \mathbf{y})$, a linear model.

## 4. PARAMETER ESTIMATION WITH AUGMENTED SPACE METHOD

In this section we propose a method that applies domain adaptation to training samples of multiple different domains to estimate parameter vector $\mathbf{w}$. First, we describe the domain adaptation method and then the algorithm used to estimate parameter vector $\mathbf{w}$.

### 4.1. Domain Adaptation with Augmented Space Method

We start with $N$ training samples of the documents that belong in the domain that we want to summarize $\{(\mathbf{x}_j^t, \mathbf{y}_j^t)\}_{j=1}^N$ as training samples of the *target domain*. We also refer to $M$ training samples whose domains are different from target domain $\{(\mathbf{x}_i^s, \mathbf{y}_i^s\}_{i=1}^M$ as training samples of the *source domain*.

To leverage training samples in the source domain for learning in the target domain, we adopt the Augmented Space Method (ASM) [5]. The method can be used independently of the learning method and is easy to implement. The method has shown efficacy in the sequential tagging problem [5].

ASM expands feature vector $\mathbf{\Phi}(\mathbf{x}, \mathbf{y})$ as follows:

$$\mathbf{\Phi^s}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{\Phi}(\mathbf{x}, \mathbf{y}), \mathbf{\Phi}(\mathbf{x}, \mathbf{y}), \mathbf{0} \rangle$$
$$\mathbf{\Phi^t}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{\Phi}(\mathbf{x}, \mathbf{y}), \mathbf{0}, \mathbf{\Phi}(\mathbf{x}, \mathbf{y}) \rangle$$

where $\mathbf{\Phi^s}$ is an expanded feature vector of the source domain examples, $\mathbf{\Phi^t}$ is an expanded feature vector of the target domain examples. The training samples in the source domain are expanded to $\mathbf{\Phi^s}$, and the training samples in the target domain are expanded to $\mathbf{\Phi^t}$. If the original feature vector has $n$ dimensions and there are training samples of $k$ source domains, the method expands the feature vector to $n \times (k + 1)$ dimensions. The expanded feature vector consists of $n$ dimensions that are shared between all domains, $n$ dimensions for one domain, and $(n - 1) \times k$ dimensions containing all zero elements. Although the above equation is for just two domains, target domain and source domain, the method can be easily expanded to the case that there are two or more source domains. We denote the expanded samples $\mathbf{\Phi^s}$ and $\mathbf{\Phi^t}$ by $\mathbf{\Phi}'$ for simplicity. We also denote the expanded parameter vector by $\mathbf{w}'$.

### 4.2. Structured Learning

In this section we explain our method to estimate parameter vector $\mathbf{w}'$. We adopt structured learning to determine the vector. Instead of learning the probabilities that indicate whether individual sentences and words are included in the summary, we learn the fitness of a summary as a set of sentences and words. We adopt the Online Passive-Aggressive Algorithm [8] to estimate parameter vector $\mathbf{w}'$ from the training samples. Since the algorithm is online, when learning parameter vector $\mathbf{w}'$ it is updated iteratively by solving the following equation:

$$\mathbf{w'}^{new} = \underset{\mathbf{w'}}{\operatorname{argmin}} \frac{1}{2} ||\mathbf{w}' - \mathbf{w'}^{old}||^2 \qquad (3)$$
$$\text{s.t. } \mathbf{w'}^\top \mathbf{\Phi}(\mathbf{x}_i, \mathbf{y}_i) - \mathbf{w'}^\top \mathbf{\Phi}(\mathbf{x}_i, \hat{\mathbf{y}}) \geq \ell(\hat{\mathbf{y}}; \mathbf{y}_i)$$

where $\mathbf{w'}^{old}$ is the parameter vector before update, $\mathbf{w'}^{new}$ is the parameter vector after update. $\ell$ is a loss function.

As the loss function, we use ROUGE [9].

$$\ell(\hat{\mathbf{y}}; \mathbf{y}_i) = 1 - \mathrm{ROUGE}(\hat{\mathbf{y}}; \mathbf{y}_i)$$

By incorporating ROUGE in the loss function, the parameter vector is strongly updated when the ROUGE score of a summary is low. Therefore, it is expected that the parameter vector will be sensitive to ROUGE score. From among the ROUGE variants, we use ROUGE-1.

When training, we use the 1-best solution to update the parameter vector. The solution is computed by the algorithm in Figure 1 as with the decoding process.

## 5. DECODING

The decoding algorithm is used when generating a summary. We show the algorithm in Figure 1.

The algorithm shown in Figure 1 was proposed by Khuller et al. [10]. Although it isn't certain that the algorithm will always find the exact solution, Khuller et al. showed that the algorithm gives a good approximate solution[1]. As shown in Figure 1, the algorithm is basically a greedy algorithm; it adds sentence $\hat{y}$ to summary $\mathbf{y}$ iteratively. The sentence $\hat{y}$ that increases the score of the summary $\mathbf{y}$ most at each iteration is added to the summary. The increase is calculated as $\mathbf{w'}^\top \mathbf{\Phi}'(\mathbf{x}_i, \mathbf{y} \cup \hat{y}) - \mathbf{w'}^\top \mathbf{\Phi}'(\mathbf{x}_i, \mathbf{y})$ and the value is normalized by the length of the sentence.

## 6. EXPERIMENT

In this section we show the efficacy of our proposed method by experiments. We use a corpus consisting of dialogues in

[1]Khuller et al. showed that the algorithm is a $(1 - 1/e)$-approximation algorithm [10].

**Table 4**. Statistics of corpus.

| Domain | Training # of samples | Training # of utterances | Training Avg. sum. rate | Test # of samples | Test # of utterances | Test Avg. sum. rate |
|---|---|---|---|---|---|---|
| FIN | 59 | 10377 | 13.38% | 60 | 8863 | 16.31% |
| ISP | 64 | 7062 | 16.29% | 59 | 9563 | 15.63% |
| LGU | 76 | 8865 | 21.20% | 56 | 7934 | 18.43% |
| MO | 70 | 9694 | 17.38% | 47 | 7305 | 20.22% |
| PC | 56 | 10088 | 13.22% | 44 | 11772 | 10.01% |
| TEL | 66 | 9774 | 16.79% | 41 | 8069 | 13.55% |

INPUT $\mathbf{w}', \mathbf{x}_i, \mathbf{\Phi}', L$
SET $\mathbf{y} = \emptyset$
SET $\mathbf{d} = \mathbf{x}_i$
WHILE $\mathbf{d} \neq \emptyset$
$\hat{y} = \underset{y \in \mathbf{d}}{\operatorname{argmax}} \frac{\mathbf{w}'^{\top} \mathbf{\Phi}'(\mathbf{x}_i, \mathbf{y} \cup y) - \mathbf{w}'^{\top} \mathbf{\Phi}'(\mathbf{x}_i, \mathbf{y})}{\operatorname{length}(y)}$
IF $\operatorname{length}(\mathbf{y} \cup \hat{y}) \leq L$ and
$\mathbf{w}'^{\top} \mathbf{\Phi}'(\mathbf{x}_i, \mathbf{y} \cup \hat{y}) - \mathbf{w}'^{\top} \mathbf{\Phi}'(\mathbf{x}_i, \mathbf{y}) \geq 0$
THEN
$\mathbf{y} = \mathbf{y} \cup \hat{y}$
ENDIF
$\mathbf{d} = \mathbf{d} \setminus \hat{y}$
ENDWHILE
$y^* = \underset{y \in \mathbf{x}_i}{\operatorname{argmax}} \left\{ \mathbf{w}'^{\top} \mathbf{\Phi}'(\mathbf{x}_i, y) : \operatorname{length}(y) \leq L \right\}$
OUTPUT $\mathbf{y} = \underset{\mathbf{y} \in \{\mathbf{y}, \{y^*\}\}}{\operatorname{argmax}} \mathbf{w}'^{\top} \mathbf{\Phi}'(\mathbf{x}_i, \mathbf{y})$

**Fig. 1**. Decoding algorithm.

**Table 3**. Dialogue Domains.

| Domain | Topic |
|---|---|
| FIN | Inquiries to banks and insurance companies. |
| ISP | Inquiries to internet service providers. |
| LGU | Inquiries to local government units. |
| MO | Inquiries to mail-order companies. |
| PC | Inquiries to computer manufacturers. |
| TEL | Inquiries to telecommunications companies. |

six domains. In this experiment, our aim is to confirm the efficacy of domain adaptation.

### 6.1. Corpus

We use contact center dialogues as the corpus. To avoid the effects caused by errors in automatic speech recognition systems, these dialogues were manually transcribed, not automatically recognized. Each dialogue consists of utterances of a customer who calls the contact center and the operator who receives the call. Since both customer and operator are Japanese native speakers, transcripts are Japanese. Each dialogue was divided into utterances. We use an utterance as the unit of summarization. Therefore, our summarizer selects a set of utterances that meets the given summary length from an input dialogue consisting of a set of utterances. The references were made by extracting utterances by annotators.

There are six domains in our corpus. We show the main topic of each domain in Table 3. As shown in Table 3, there are various topics in dialogues. Our aim is to summarize each domain by leveraging the training samples of the target and other domains.

We show the statistics of our corpus in Table 4. # of samples is the number of samples contained in training or test set of each domain, # of utterances is the number of utterances contained in training or test set of each domain. Avg. sum. rate is the average summarization rate that is calculated by dividing the reference length by the average length of original input documents. The number of references is 250 characters in both training and test set.

### 6.2. Setting

Following the previous work on domain adaptation [5], we also compare the following four methods.

1. **Source domain only**. When learning, a learner uses only training samples whose domains are different from the domain of documents to be summarized. When test examples in FIN domain are summarized, the parameter vector is trained using ISP, LGU, MO, PC and TEL domains.

2. **Target domain only**. When learning, for each domain, the learner uses only training samples whose domain is the same as the domain of documents to be summarized. When test examples in FIN domain are summarized, the parameter vector is trained using training samples in FIN domain. Hence this situation is the same as usual supervised learning.

3. **All domains**. When learning, a learner uses training samples of all domains without distinction. When test examples in FIN domain are summarized, the parameter vector is trained using training examples in FIN, ISP, LGU, MO, PC and TEL domains.

4. **Domain Adaptation**. Proposed method. We adopt the domain adaptation method mentioned above.

We call these methods Method (1)-(4), respectively. The result we want to clarify is whether our proposed approach, Method (4), is superior to usual supervised approach, Method (2). If this is true, contact center dialogue summarization systems should use a domain adaptation technique.

We used ROUGE-1 [9] to evaluate our method.

### 6.3. Results and Discussions

We show a result of our experiment in Table 5. Values in each row and column are the values of ROUGE score for the corresponding methods and domains, respectively. Bold-faced values are the highest scores in each domain.

As shown in Table 5, our proposed method achieves the best score in three of six domains, FIN, ISP and MO. Method

**Table 5**. ROUGE-1 results.

| Domain | Method (1) | Method (2) | Method (3) | Method (4) |
|---|---|---|---|---|
| FIN | 0.437 | 0.558 | 0.551 | **0.584** * |
| ISP | 0.465 | 0.525 | 0.508 | **0.543** |
| LGU | 0.466 | **0.514** | 0.506 | 0.500 |
| MO | 0.563 | 0.632 | 0.601 | **0.635** |
| PC | 0.215 | **0.394** | 0.330 | 0.367 |
| TEL | 0.421 | 0.390 | **0.457** | 0.418 |

(4) is superior to the usual supervised approach, Method (2), in four of six domains. In FIN domain, Method (4) score surpassed Method (2) by a statistically significant margin at the 95% level.

In FIN domain, Method (4), achieved the best result, followed by Method (2), Method (3) and Method (1). This trend is also observed in ISP and MO domains. In these domains, supervised learning method, Method (2), achieves good results, though the domain adaptation method leveraged the training samples in different domains to improve usual supervised learning.

In LGU and PC domains, Method (2) surpassed Method (4). This result suggests that their source samples have no training samples that are similar to and that are useful as samples for the target domain. Actually, dialogues in PC domain are particularly troublesome, because customers frequently raise technical problems about their computers and hence the dialogues often become long, as its Avg. Sum. Rate shows. In such long dialogues, important utterances are at the end of dialogues, while in other domains important utterances are at the front.

In TEL domain, Method (2) failed to match Method (1). That is, training samples in source domains are more useful than training samples in the target, TEL domain. This result implies that in the TEL domain there is some kind of gap between training and test examples. Previous work [5] pointed out that if a method that leverages only samples in a source domain beats a method that use only samples in the target domain, domain adaptation is not effective. This result confirms that conclusion.

## 7. CONCLUSION

In this paper we proposed a method to improve the quality of extractive summarization by making use of training examples whose domains are different from that of the target domain. We adopted the Augmented Space Method to adopt samples from source domains to the target domain and validated its efficacy by experiments. By our experiments, our proposed domain adaptation approach achieved the best results.

An immediate research direction is to leverage training samples in other than contact center dialogues, such as news documents. There are a lot of training samples in news domains, hence using them is promising.


### Acknowledgments

We would like to thank the anonymous reviewers for their comments.


## 8. REFERENCES


[1] Roy J. Byrd, Mary S. Neff, Wilfried Teiken, Youngja Park, Keh-Shin F. Cheng, Stephen C. Gates, and Karthik Visweswariah, "Semi-automated logging of contact center telephone calls," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM)*, 2008.

[2] Ryuichiro Higashinaka, Yasuhiro Minami, Hitoshi Nishikawa, Kohji Dohsaka, Toyomi Meguro, Satoshi Takahashi, and Genichiro Kikui, "Learning to model domain-specific utterance sequences for extractive summarization of contact center dialogues," in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling)*, 2010.

[3] Ryuichiro Higashinaka, Yasuhiro Minami, Hitoshi Nishikawa, Kohji Dohsaka, Toyomi Meguro, Satoshi Kobashikawa, Hirokazu Masataki, Osamu Yoshioka, Satoshi Takahashi, and Genichiro Kikui, "Improving hmm-based extractive summarization for multi-domain contact center dialogues," in *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT)*, 2010.

[4] Tsutomu Hirao, Hideki Isozaki, Eisaku Maeda, and Yuji Matsumoto, "Extracting important sentences with support vector machines," in *Proceedings of the 19th international conference on Computational linguistics (Coling)*, 2002.

[5] Hal Daume, III, "Frustratingly easy domain adaptation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2006.

[6] Oana Sandu, Giuseppe Carenini, Gabriel Murray, and Raymond Ng, "Domain adaptation to summarize human conversations," in *Proceedings of ACL Workshop on Domain Adaptation in NLP*, 2010.

[7] Shasha Xie, Benoit Favre, Dilek Hakkani-Tur, and Yang Liu, "Leveraging sentence weights in a concept-based optimization framework for extractive meeting summarization," in *Proceedings of Interspeech*, 2009.

[8] Crammer Koby, "Online passive-aggressive algorithms," *Journal of Machine Learning Research*, vol. 7, no. Mar, pp. 551–585, 2006.

[9] Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of ACL Workshop Text Summarization Branches Out*, 2004, pp. 74–81.

[10] Samir Khuller, Anna Moss, and Joseph Naor, "The budgeted maximum coverage problem," *Information Processing Letters*, vol. 70, no. 1, pp. 39–45, 1999.