Engineering and Applied Sciences
2021; 6(5): 82-85
http://www.sciencepublishinggroup.com/j/eas
doi: 10.11648/j.eas.20210605.11
ISSN: 2575-2022 (Print); ISSN: 2575-1468 (Online)

Tscan: Dialog Structure Discovery Using Scan, Adaptation of Scan to Text Data

Apurba Nath, Aayush Kubba

Voicezen India Pvt Ltd, Gurugram, India

Email address:
apurba@voicezen.ai (A. Nath), aayush@voicezen.ai (A. Kubba)

To cite this article:
Apurba Nath, Aayush Kubba. Tscan: Dialog Structure Discovery Using Scan, Adaptation of Scan to Text Data. *Engineering and Applied Sciences.* Vol. 6, No. 5, 2021, pp. 82-85. doi: 10.11648/j.eas.20210605.11

Received: August 1, 2021; **Accepted**: August 18, 2021; **Published**: September 7, 2021

Abstract: Can we learn dialog structure from existing dialogs without ontology or domain assumptions. Understanding dialog structures from existing task oriented human human dialogs can help us automate these dialogues in a better way. Traditionally dialog structures have been created using ontologies that are created by domain experts. However, in our experience getting the ontology right is difficult and time consuming. Like other such tasks an unsupervised approach may do better than hand crafted rules. We propose an unsupervised dialog structure discovery approach that is based on SCAN (Semantic Clustering using Nearest Neighbors). Our approach comprises of two steps, the first being creating clusters of utterances and the second being creation of a structure using inter-cluster transition probabilities. Our main contribution in this paper is the adaptation of SCAN on text data. Unlike the SCAN approach for images, for text we did not train a separate pretext model and were able to use BERT for the same. Similarly for neigbor discovery, instead of augmentation we were able to leverage data variety. Evaluation metrics on dialog structures are a bit subjective, so we have used statistical measures as proxies for structure quality. We have also included our results on an internal human human task oriented 100k dialog dataset. We think SCAN like approaches are very promising for problems that use embedding similarities and should be further explored.

Keywords: Unsupervised Learning, Dialog Structure Discovery, Text Classification, Clustering

1. Introduction and Prior Work

Dialog structure discovery is an important problem given the increased efforts in automation of text and voice response systems. Unlike the simulated dialogs or human bot interactions, human human interactions are richer (larger vocab and variation) and have more number of turns. For example typical dialog datasets created from SimDial have less than 1.5k vocab spread with 20 n-grams covering majority of generated responses. Compared to this our internal human human task-oriented dialog includes over 5k of common closed (proper nouns excluded) vocab with the most common 20 n-grams failing to cover even 1% of utterances.

Many approaches rely on generational models which are trained on the dialog data e.g VRNN approach (Shi, Zhao) [2] or DVAE-GNN (Xu, Che) [3], some are schema-driven [11, 12, 14] and others are based on data driven belief tracking [13]. There are also approaches using transformers like BERT models [15]. In our experience BERT when trained on large in-domain data captures semantic information abundantly, we can cluster on these embeddings but balance and interpretability of these clusters is still a challenge.

On image classification without labels SCAN [1] has achieved great results. Their approach comprises of obtaining semantically meaningful features, learning a clustering approach and then self-labelling for interpretable clusters. They use image transforms and nearest neighbors in this work. They use confidence and consistency both as part of their objective function while training the clustering model which creates balanced clusters. It is also not negatively impacted by overclustering.

2. Our Approach

Each dialog is made of T turns (A_1, U_1), (A_2, U_2),..., (A_t, U_t) where A_t is the agent utterance at t-th dialogue turn and U_t the user utterance. These dialogs are task oriented and may have multiple exchanges (multiple tasks) in the same dialog.

Our goal is to eventually find out any correlation between A_t and U_t and U_t and A_{t+1}. We try to first reduce the size of the space (because of vocab variety) by assigning the utterances to

clusters. With a 20 state cluster, now this becomes a problem of matching the clusters among each other. For example, assuming A_t belongs to agent cluster AC_0 and user utterance U_t belongs to UC_3 we can group them with any other turn which similarly have AC_0 and UC_3. We create transition probabilities between the cluster combinations (AC_i, UC_j), these transition probabilities are then used for dialog states.

A simplified version of these steps are:

1. Use in-domain trained BERT for semantic embeddings.
2. Train SCAN model with nearest neighbors on 10k A_* agent utterances and U_* user utterances.
3. Create clusters using this model and apply self-labels on it.
4. For each Dialog turn assign the agent cluster and customer cluster.
5. Create a transition map between agent and customer turn and customer to next agent turn.
6. Create dialog flows with these transition states, each cluster is represented by it's equivalent label.

2.1. Models

The BERT model is used as pretext task to satisfy the equation (1) of SCAN paper, replicated here for convenience

$$\min_{\theta} d(\varphi_{\theta}(X_i), \varphi_{\theta}(T[X_i])) \quad (1)$$

fine tuning or training on large volume of in-domain data helps us create such a model which meets the goals of other pretext tasks like the ones mentioned in [5-8] Any MLM evaluation task can be used to check the semantic quality of the model. The bert model pretext results are in alignment with.

The SCAN model needs to satisfy the equation (2) of the SCAN paper, a simplified form of that equation is

$$\text{loss} = \text{consistency\_loss} - \text{entropy\_weight} * \text{entropy\_loss} \quad (2)$$

consistency loss is BCE between anchors and neighbors while entropy loss is mean of anchors probability. The entropy constituent helps in balancing the distributions within the clusters.

2.2. Our Experiences

Unlike original SCAN implementation we do not use transformations or augmentation, instead we rely on the variety of data to provide the relevant neighbors. We also do not build a pretext model but use a BERT model trained on in-domain data for the same. Our experiments show that in spite of these deviations from the SCAN approach we are able to create a interpretable dialog structure from the balanced well defined clusters created by SCAN. We use two statistical measures as evaluation metrics to understand the cluster quality and our experiments show that TSCAN (text SCAN) does better than K-means on both these measures.

2.3. Evaluation Metrics

The goal of clustering is to evenly balance the utterances between the clusters. This means we should not have any cluster that is too big. To compute the distribution score, we use.

Distribution we want the clusters to be balanced, that means each cluster should have nearly the same number of members. A good measure of the same is

$$\sum x log(x) \quad (3)$$

where x is the ratio of members vs total elements. Though this number is not comparable across cluster sizes, within a cluster size it is a good indicator of the distribution. For comparison across cluster sizes we can use deviation from ideal distribution.

Confidence We expect similar utterances to end up in the same cluster. As we already have some pre-trained intent models, we can check that utterances with the same intent end up together. We want the number of clusters to be as low as possible. A good measure of togetherness is the mean and standard deviation of cluster membership. For example, in case we have a greeting intent, we would want all the greeting intents to end up in the same cluster. A scenario where it is spread between 3 different clusters out of 20 is better than where it is spread between 8. Mean and standard deviation of these two scenarios give a good indication of the distribution.

3. Experimental Setup

3.1. Datasets

We use an internal dataset containing 100k human human dialogs. These dialogs are task oriented but may have multiple tasks in the same dialog. The average number of turns is 15 and the closed vocab is greater than 5k. This are real contact center calls, so the start and the end are closer to a scripted pattern, while the middle of the conversation is driven by the customer. In our prior attempts at manual annotation of the same using DIT++ annotation scheme [4] all 11 typical types of utterances were present in this dataset.

3.2. Training

We used a BERT model trained on similar dialogs for our pretext task. We used faiss based vector similarity for closest neighbors. This was trained using SCANloss. Once the clustering model was trained, it was applied on the dataset and utterance clusters were created.

We then computed transitions from one utterance cluster to another. We picked up all major transitions (above transition probability of 1/k) and create a dialog structure based on this. We also provided labels to each cluster based on the self-labelling work of SCAN. It is similar to the pseudo-labelling approaches discussed in papers like [9, 10].

4. Results

As a proxy for clustering quality, we used consistency and confidence measures and compared these values with a K-means approach.

4.1. Consistency

For our approach a consistent outcome is when all clusters are equal in size. To measure this, we use sum of nlogn where n is membership ratio. In an ideal case the utterances will be evenly distributed among the k clusters giving the highest nlogn score. For example, if we have 20 clusters, the best case is where each cluster has 0.05%, this has the highest score. On the other hand, if some clusters have 0.01 and others 0.1 it will have a lower score.

For the clustering approach, for a 20 cluster SCAN vs K-means approach, K-means shows a distribution score of -2.64, SCAN achieves -2.77 while an ideal distribution is -2.995.

4.2. Confidence

Utterances with the same intent should be clustered together. As we know some of the intents in this dialog set, we can check if those utterances are classified into the same cluster. For example, all greetings should go into same or similar clusters, same for payment inquiry.

For two intents greeting and payment_inquiry, the results were:

intent: payment_inquiry with K-means

nobs=8, minmax=(6, 43), mean=12.12, variance=162.98, skewness=2.07, kurtosis=2.63

with SCAN

nobs=6, minmax=(6, 60), mean=16.16, variance=468.97, skewness=1.73, kurtosis=1.07

intent: greeting with K-means

nobs=6, minmax=(0, 91), mean=16.17, variance=1347.77, skewness=1.78, kurtosis=1.18

with SCAN

nobs=5, minmax=(0, 95), mean=19.4, variance=1786.30, skewness=1.50, kurtosis=0.25.

K-means utterances of the same intent payment_inquiry are spread over 8 clusters with 43% of them being in one while with SCAN, it is present in only 6 of the clusters with 60% of them located in a single cluster.

4.3. Transitions

These clusters were then used to map transition probabilities, all transitions with probability less than 0.06 were ignored. In Figure 1, the clusters are represented by alphabets a to t, start and end are represented as by ^ and $ symbols.

As we can see in figure 1, the dialog starts from ^ to n or o. n and o are self-introduction (I am x) and rpc-inquiry (am I talking to y). From n and o it goes to other nodes including d which is brand-intro (I am calling from brand z). There are nodes like e, j, r which are common utterances like e (the reason why I called..), j (please hold for a minute) and r (your transaction id is) which can be reached from most of the other nodes. r and g (when will it be done) are the most common termination nodes. Given below in the table are few of the node names along with the self-labels. For our understanding a manual label column has also been added. These self-labels are utterances from the actual dialogs which were mined using the self-labelling method of SCAN.

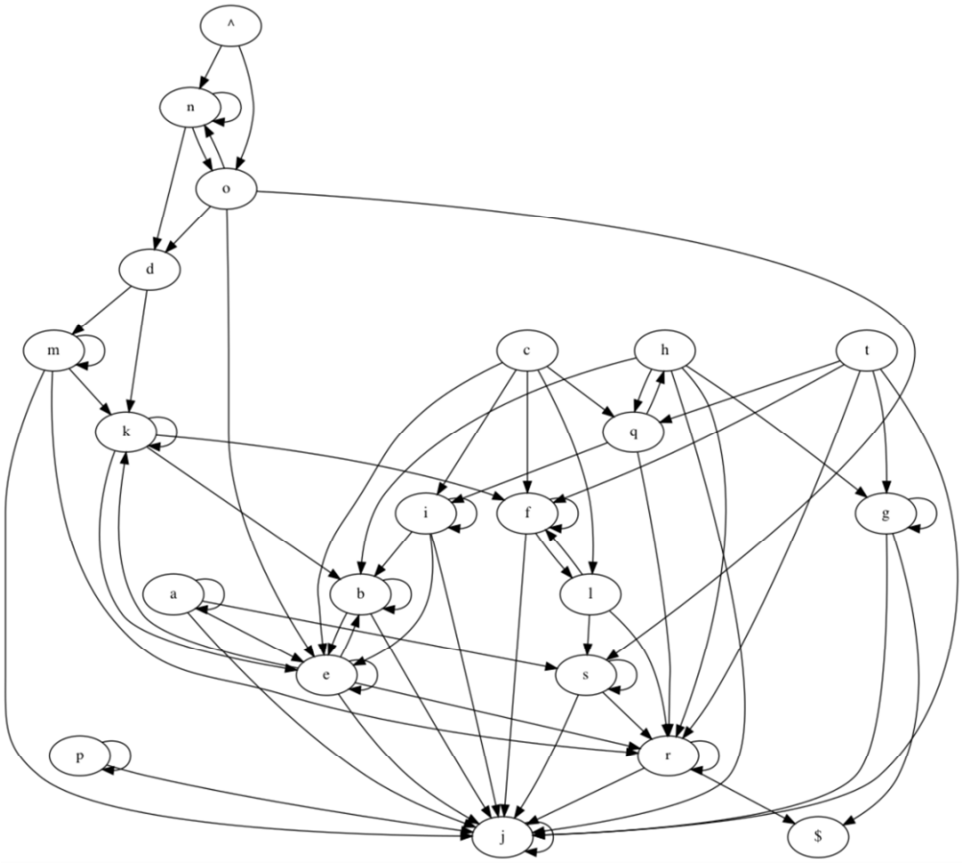

Figure 1. *Dialog-Structure.*

Table 1. *Utterances from the actual dialogs which were mined using the self-labelling method of SCAN.*

cluster-name	manual-label	self-label
^	start-node	Start Node
n	self-introduction	I am X
o	rpc-inquiry	Am I talking to Y
d	brand-intro	I am calling from brand Z
m	product-info	This call is about product Z that you purchased last month
k	payment-inquiry	Have you made the payment for the last installment
f	amount-info	The amount is five thousand three hundred dollars
l	date-reminder	Your due date is third of June
r	number-intimation	Your transaction id is five nine eight zero double two
g	payment-date-inquiry	When will the payment be done
$	end-node	End Node

5. Conclusions

The two-step process introduced by SCAN for image clustering can be applied for text too. The idea of creating a model that gives a confidence score for cluster membership and balances the cluster distribution may be very useful in all embedding based clustering approaches. While working with embedding one of the biggest challenges has been the problem in interpretation of distances. At times embeddings with a distance of 300 is good while at others 100 is also not good, with SCAN we are able to convert the distance into a cluster membership probability score free from all distance subjectivities. Moreover, with neighbor selection, we can actually influence how automated clustering will be done. We can use these clustering attributes for interpretable structure discovery on various other problems. We hope that the dialog structure discovery approach helps people understand that for real world structures which are diverse and complex, approaches like this stand a better chance than hand crafted schema-based approaches.

References

[1] Van Gansbeke, Wouter and Vandenhende, Simon and Georgoulis, Stamatios and Proesmans, Marc and Van Gool, Luc. Scan: Learning to classify images without labels. In: Proceedings of the European Conference on Computer Vision (2020).

[2] Qiu, Liang and Zhao, Yizhou and Shi, Weiyan and Liang, Yuan and Shi, Feng and Yuan, Tao and Yu, Zhou and Zhu, Song-Chun. Structured Attention for Unsupervised Dialogue Structure Induction. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1889–1899. (2020).

[3] Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, Ting Liu. Discovering Dialog Structure Graph for Open-Domain Dialog Generation. arXiv preprint arXiv: 2012.15543, (2020).

[4] Harry Bunt. 2009. The dit++ taxonomy for functional dialogue markup. In AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts, pages 13–24.

[5] Wu, Z., Xiong, Y., Yu, S. X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR (2018).

[6] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: arXiv preprint arXiv: 1911.05722 (2020).

[7] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv: 2002.05709 (2020).

[8] Misra, I., van der Maaten, L.: Self-supervised learning of pretext-invariant representations. In: CVPR (2020).

[9] Sohn, K., Berthelot, D., Li, C. L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv: 2001.07685 (2020).

[10] Asano, Y. M., Rupprecht, C., Vedaldi, A.: Self-labelling via simultaneous clustering and representation learning. In: ICLR (2020).

[11] Su Zhu, Jieyu Li, Lu Chen, and Kai Yu. 2020. Efficient context and schema fusion networks for multi-domain dialogue state tracking. arXiv: 2004.03386.

[12] Lu Chen, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. 2020. Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In AAAI, pages 7521–7528.

[13] Nikola Mrkšic´, Diarmuid O´ Se´aghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In Proceedings of the 55th Annual Meeting of the ACL (Volume 1: Long Papers), pages 1777–1788.

[14] Liliang Ren, Jianmo Ni, and Julian McAuley. 2019. Scalable and accurate dialogue state tracking via hierarchical sequence generation. arXiv preprint arXiv: 1909.00754.

[15] Yan Zeng, Jian-Yun Ne, Multi-Domain Dialogue State Tracking – A Purely Transformer-Based Generative Approach, In: arXiv preprint arXiv: 2010.14061 (2020).