Variational Autoencoding Dialogue Sub-Structures Using a Novel Hierarchical Annotation Schema *

1st Maitreyee Tewari
Dept. of Computing Science
Umeå University
Umeå, Sweden
maittewa@cs.umu.se

2nd Michele Persiani
Dept. of Computing Science
Umeå University
Umeå, Sweden
michelep@cs.umu.se

***Abstract*—This work presents a novel method to extract sub-structures in dialogues for the following genres: human-human task driven, human-human chit-chat, human-machine task driven, and human-machine chit-chat dialogues. The model consists of a novel semi-supervised annotation schema of syntactic features, communicative functions, dialogue policy, sequence expansion and sender information. These labels are then transformed into tuples of three, four and five segments, the tuples are used as features and modelled to learn sub-structures in above mentioned genres of dialogues with sequence-to-sequence variational autoencoders. The results analyse the latent space of generic sub-structures decomposed by PCA and ICA, showing an increase in silhouette scores for clustering of the latent space.**

***Index Terms*—Variational autoencoders, Attention Layer, Dialogue sub-structures, Conversation Analysis, Dialogue control functions, Dialogue Policies, LSTM, Hierarchical Agglomerative Clustering**

I. Introduction

Commercialisation of technology (virtual and physical) capable of interacting in natural language such as chat-bots and home assistants [1], [2] has accelerated the research domain of dialogue management systems (DMS). In order for these DMS systems to establish and succeed in the human society, we need to start advancing towards building strategies that are domain agnostic such that DMSs can manage versitality (for instance sudden changes of topics) and complexity (introduced by errors and miscommunications) that human participant might introduce while interacting with a machine.

This work assumes that dialogue sub-structures can partially provide a solution for DMS to managing change of topics and miscommunications. Hence, the over-arching goal of our work is to build domain agnostic strategies for DMS systems capable of managing topic change and miscommunication. To this end, exploration of dialogue sub-structures for English language that can be generic for multiple domains (currently the focus is on task-driven/chit-chat human-human and task-driven/chit-chat human-machine dialogues) has been performed.

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 721619 for the SOCRATES project.

So far the research community has explored structure in dialogues for specific domains [3]–[5] or with a specific schema [6], [7] relating to a research objective. In this work we want to investigate if we can decode dialogue sub-structures that are generic for multiple domains and are longer than two turns. The above question is investigated and we answer it in two stages: (i) by proposing a novel generic schema for annotating dialogues from four different genres and (ii) training a sequence to sequence variational autoencoder model (Seq2SeqVAE) that decodes dialogue sub-structures for $3-5$ turns.

The main contribution of this work is two-folded: *(i)* a generic and flexible annotation schema allowing the annotation of dialogue corpus for diverse genres. *(ii)* a Seq2SeqVAE model with attention layer is provided for generating generic and long sequences of dialogue sub-structures.

The rest of the article is structured as follows: Section II provides related work and necessary background. Followed by the method Section III, highlighting the approach on the proposed research gap on modelling generic dialogue sub-structuring. Section IV presents the statistical and empirical results of Seq2SeqVAE. Section V concludes the article with some discussion and future work.

II. Related Work and Background

Conforming to the ideas of [6], [8], we believe generic sub-structures in dialogues can be utilised for building DMS strategies capable of managing dialogue complexity (long distance pronoun resolution, topic shifts, and multiple communicative functions introduced with-in a single turn) of human participants.

Interest in exploring dialogue structures can be traced back to [3], where the authors proposed to learn tree like intention structure in task-driven dialogues, using *VERBMOBIL* corpus [9]. A grammar induction algorithm was proposed in combination with plan recognition for parsing the annotated context-free grammar trees. Authors in [10] explored improvements of adding the last dialogue act of a participant to reinforcement learning based DM strategies. They implemented and tested the system for three flight booking systems. The authors in [4], presented a structural analysis of task driven dialogues with dialogue acts, topic and a task type. They

conducted experiments on sub-structure of dialogues with chunk-based, parse-based and hierarchical methods. The work in [7] models expectation in DMS by structuring dialogues using adjacency pairs [11]. In [6], the authors proposed a two step methodology of extracting two dimensional sub-structures in dialogues, followed by clustering them. Authors in [8] demonstrated the significance of dialogue sub-structuring for building domain agnostic dialogue models using CA. They explored sequence expansion and developed an annotation tool with CA and dialogue control functions to annotate dialogues. Our work is in close line with [12], [13], however, is more directed to exploration of computational models for conversational analysis (CA) as in [8], [14] and using an unsupervised method as in [6]. This work is novel in at-least two aspects; first we use a semi-supervised annotation schema for CA with $tri, four, five$-grams compared to manual labelled uni-grams or 'raw' bi-grams as input in the previous works. Other contribution is a generic schema for dialogue annotation and Seq2SeqVAE with higher generalising capability compared to specific annotation schema, and clustering method in [6].

The sociological domain of conversation analysis (CA) [15] provides sub-structuring schema directly applicable to dialogues. The primary assumption in CA is that dialogues have sequential organisation and a brief over-view is provided next.

A. Sequence Expansion (SE)

Sequence expansion (SE) [16] in CA focuses on the placement of an utterance, within a dialogue, in order to understand the meaning and significance of the performed action relevant to the dialogue. According to conversation analysis (CA) [15], dialogues are a composition of units produced sequentially, using turn construction units (TCU) of verbal and non-verbal cues with reflexive relationship between the prior and the following TCUs (for simplicity this work uses utterances and TCUs synonmously). The base form of TCUs are formally known as adjacency pairs [11] with first pair part (FPP_{base}) and second pair part (SPP_{base}). Instances of commonly found adjacency pairs are greeting-greeting, request-accept/reject, offer-accept/reject, question-answer. The characteristics of adjacency pairs; is to be composed of two turns, produced by different speakers, and relatively ordered to form a pair. SE relaxes the order of adjacency pairs by have a preceding *(FPP_{pre}, SPP_{pre})*, intervening *(FPP_{insert}, SPP_{insert})* and/or a follow expansion *(FPP_{post}, SPP_{post})*

Table I provides a sample of a chit-chat dialogue from our corpus, labelled with SE and dialogue control functions capturing the multi-functional and multi-turn nature of dialogue sub-structures. In Table I the first part of adjacent pair denoted by FPP_{base} was not followed by its second pair, then the dialogue continues with Human2 using multiple communicative functions FPP_{insert} to understand the real question by Human1, to which Human1 in the second last instance provides a positive feedback together with a re-formed question FPP_{base} which is addressed finally by Human2 with an answer SPP_{base}. The next sections provide details about the method employed in this work and the novel annotation schema.

III. Methodology

Our method consists of following steps 1) corpus creation 2) data pre-processing 3) data annotation 4) training the annotated data on the baseline models, and 5) training the annotated data on variational autoencoder models.

A. Corpus

Our corpus consists of a total of 89 dialogues and 5268 utterances, of which 44 were synthetically created dialogues between a human (H) and an agent (A) [1]. The agent A is assumed to be either a physical robot or a virtual assistant. The corpus consists of the union of four main dialogue sources: *synthetic dialogues*, *DialogBank*, *DBDC3*, and *transcribed dialogues* illustrated later in the section.

Synthetic dialogues are hand-crafted dialogues contextualized in an envisioned scenario where A is situated inside the house of H as a companion robot assisting in tasks. These tasks are meal cooking, calendar reminders such as calls, visitors, medicine intake, physical exercise, yoga sessions, and cognitive exercise sessions, playing board games, giving instructions for computer or physical games, ordering food online, assisting with garbage disposal, helping select a suitable garment online, and casual talks such as choosing a hair-cut or why to do physical exercise.

From **DialogBank** [17]15 samples were selected, this corpus came already annotated by gold-standard ISE 24617-2 schema. Out of 15, there were 3 task-driven dialogues from MapTask [19], 4 chit-chat dialogues from Switchboard [20], 5 task-driven dialogues between human and a conversational agent from D-Box [2] and 3 dialogues from Trains corpus [21].

DBDC3: 28 dialogues were selected from dialogue breakdown detection challenge (DBDC3) [18], consisting of both chit-chat and task-driven conversations between a human and a chat-bot.

Transcribed: the corpus also consists of 2 transcribed dialogues between one of the researcher and the two colleagues.

We further categorised the dialogues from these different sources into four genres considered: 3 Human-Human task-driven dialogues from Map Task; 6 Human-Human chit-chat transcribed dialogues from Switchboard corpus; 46 Human-machine task-driven dialogues from DBDC3, Trains, D-Box, and Synthetic corpus; 34 Human-machine chit-chat dialogues from DBDC3 and Synthetic corpus. Even though the corpus may seem biased towards human-machine dialogues however, the sample of utterances for each genre were nearly balanced because average size of human-human dialogues was ≈ 500 turns compared to human-machine with number of turns ≈ 40.

[1]The authors are well aware of the limitation that a small corpus imposes on deep learning models in general. However, we have made sure that the corpus is diverse enough and contains samples from publicly available sources (DialogBank [17], DBDC3 [18]) and real world (transcribed dialogues mentioned later). The statistical results and the manual analysis indicated as you would be reading later in the article that the model's performance was significantly good even though the training data was comparatively small.

[2]https://www.idiap.ch/project/d-box

Utterance Segments	Dialogue Control Functions, SE
Human1: what are you most known for in history of humankind?	Question, FPP_{base}
if its not too general for a question	Inform, FPP_{insert}
Human2: that is a very good question	Positive Feedback, SPP_{insert}
i m not sure if i can answer it	Inform, FPP_{insert}
give me just a second	Turnkeep, FPP_{insert}
not sure that i can fully answer that question	Inform, FPP_{insert}
Human1: okay	Positive Feedback, SPP_{insert}
can you name one thing that many people know that you did	Question, FPP_{base}
Human2: mh	Turnaccept, FPP_{insert}
i was married to two kings and i participated in the second crusade while queen of France.	Answer, SPP_{base}

TABLE I: In the dialogues from our corpus, highlighted that questions are answered differently for instance with a confirmation or answered way later, or are preceded or succeeded by other dialogue control functions as illustrated in above example.

B. Labelling Method

Among the main contributions of this work is a generic sub-structuring schema presented in Figure 1 that uses specific syntactic features, dialogue control functions and policies, and sequence expansion.

The motivation behind creating this schema came from the knowledge gap that the authors identified in the dialogue community, regarding the lack of standardisation of annotation schema and that can be generic enough. This work instead claims that our proposed schema can be consistently used for any domain and dialogue scenario (human-human or human-machine.)

We assume, based on previous works [22], [23] that dialogues are not just spoken words but they have dimensions including, but not limited to syntax, semantics, communicative, emotive, organisational, and intention. The abstraction level of these dimensions is also at incremental levels. For example a dialogue consists of turns, guided by communicative actions and some other conversational norms. Each turn has one-to-many relationship with utterances, which can further be divided into lexical and non-lexical items. Hence, from a single lexical or non-lexical item we can only abstract the syntax, many such syntax values make up a segment. From such segments we can abstract syntactic relations and the communicative action they perform on behalf of the dialogue participant (also, semantics can be extracted at this stage). An organisation of such segments can be used to abstract the policies they might be using and the sort of sequences they form. Since, the abstraction of different dimensions needs to be performed at different levels, a hierarchical schema was most suitable, as also indicated in [24].

Next, pre-processing of the corpus was performed by removing punctuation marks and special symbols, followed by the segmentation of each utterance in a dialogue following ISE24617-2 [17] to units called *segments* i.e. the longest stretch of verbal (but, okay, yes, fine, why, what) or non-verbal cues (laughing, hmm, umm.)

At the first stage of the labelling procedure (Figure 1), each utterance is decomposed into segments. For example, each line in Table I is a segment and has an associated dialogue control function. At the second stage of labelling, we label each segment with dialogue control functions and extract syntactic features, such as through dependency parsing. Since some syntactic features are very sparse in dialogue segments, we filtered words based on their dependency or POS tags: *(interjections)* such as okay, oh, hmm, *(adverbs)* such as why and how and *(nouns)* such as what, and graphs of *(subject-object-verb)* and *auxiliary verb*.

At the third stage dialogue control functions are present together with the extracted syntactic elements and graphs. These combined dialogue features are then labelled with one of the categories of sequence expansion (SE) as explained in Section II and novel dialogue policies that we proposed in this work. We claim that SE components can allow the connection between dialogue control functions, which results into longer sequences.

At the same level of SE, this work uses *dialogue policies* [25], that can allow to understand the nature of the dialogue control functions in terms of the demand they pose on the addressee. The three dialogue policies are: *(i)* co-occurring policy: where dialogue control functions for example turn taking, stalling, inform, and correction occur together with a question, suggest, or request without requiring an addressal. *(ii)* Binding policy: when a dialogue control function binds the addressee to respond with another specific dialogue control function, for instance questions, greetings, request, offer, instruct bind the address to answer them appropriately. *(iii)* Progressive policy: when a dialogue control function allows the addressee to change the topic, or even end the dialogue, for instance in case of inform, answer, accept, confirm, agree, disagree etc.

The above explained features together with sender information are used as sequences($tri, four, five$-grams) with a variational autoencoder model to evaluate whether such models are capable of generating generic long sequences representing dialogue sub-structures. We assume dialogue control functions as the **baseline schema** with which our schema could be compared with. The experiments ran on different selection of labels and also on the entire schema.

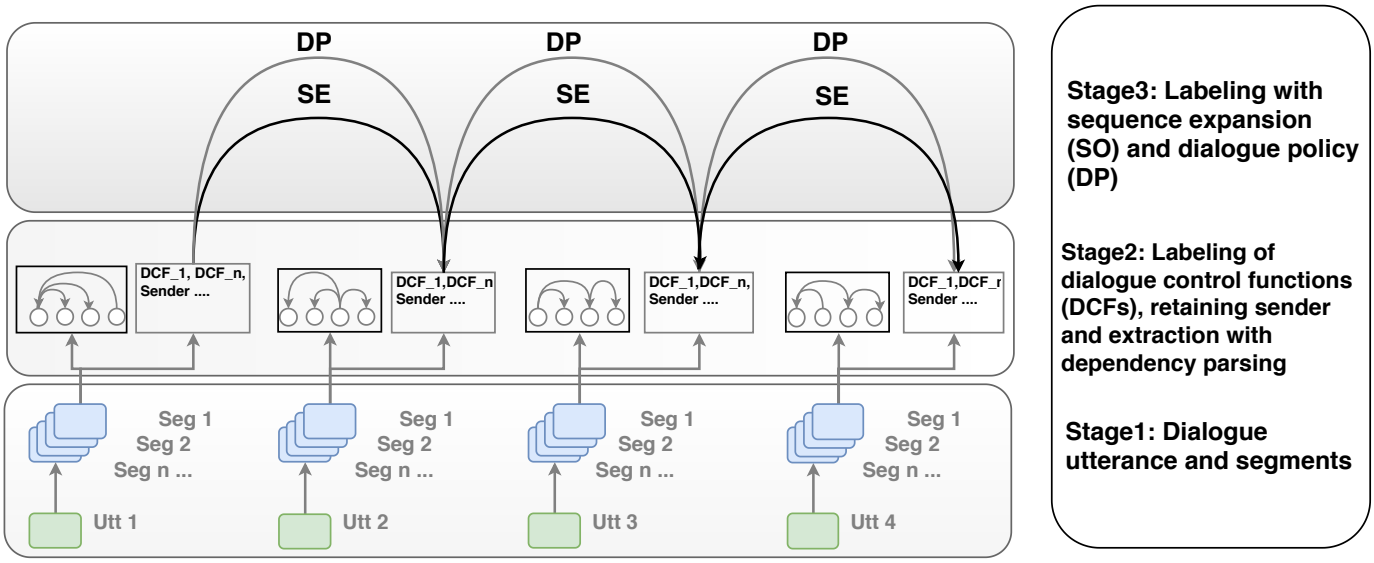


Fig. 1: Three different stages of labelling of sequences: segmenting dialogue utterances, syntactic parsing and annotation of segments and then establishing sequence expansion and dialogue policy relations. Here, $\mathbf{Utt1}, \mathbf{Utt2}, ..., \mathbf{Uttn}$ represent utterances of a dialogue, $\mathbf{Seg1}, \mathbf{Seg2}, ..., \mathbf{Segn}$ are the segments of utterances, $\mathbf{CF}$ are dialogue control function associated with each segment, graph represents dependency parse of a segment, $\mathbf{SE}$ denotes the sequence expansion label, and $\mathbf{DP}$ are the dialogue policies between the dialogue control functions of adjacent utterances

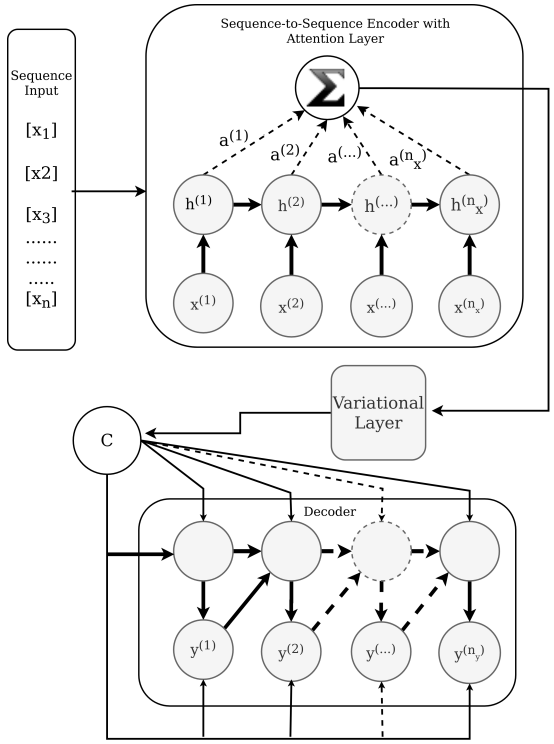


Fig. 2: Architecture of the Seq2SeqVAE with attention layer.

C. Variational Autoencoders (VAE)

Autoencoders (AE): [26] are a family of unsupervised learning methods used for learning corpus representation for transfer learning [27], and sequence to sequence learning. In essence AE copies its input to the output with minimum distortion. Its hidden layer $\mathbf{h}$ is an encoder function $\mathbf{h} = f(x)$ and the output is a decoder function $r = g(h)$. Seq2Seq autoencoder (Seq2SeqAE) allows the mapping of a fixed sequence of input with fixed sequence of output, with variation in input-output length. It can be constructed with three components: an *encoder*, *intermediate* encoder and *decoder*.

This work uses a varient of AE, namely sequence to sequence variational autoencoders (Seq2SeqVAE). It comes with a regulariser that reduces the chances of over-fitting during the training phase, where it encodes the input sequence as a distribution over the latent space, and the decoder samples a single point from it. The sampled point is decoded, reconstruction error is computed and then back-propagated to the network. Our Seq2SeqVAE is complemented with an attention layer to build the context during encoding of the sequence data and then providing that context to the decoder enabling it to focus on specific aspects of the encoded data while generation. The Seq2SeqVAE architecture used for this work is depicted in Figure 2, where sequence of features from our annotation schema is used as input to the model, encoded with the LSTM layers. The attention layer starts working with the encodings at time $t+1$ and accumulates the important information from LSTM layer, the variational layer create a distribution of the encoding, finally the decoder selects the input and generates the output with sequences.

D. Baseline Models

As baseline we use hierarchical agglomerative clustering (HAC) and a basic sequence to sequence autoencoders (Seq2SeqAE). A Seq2SeqAE is an encoder with a stack of

LSTM layers [28] with units processing and aggregating the required information one sub-unit of the input at a time and then forwarding it ahead. The decoder consists of stack of three layers: the first intermediate layer that stores all the information from the encoder, and the next two LSTM layers to predict the output using the Softmax function.

Hierarchical agglomerative clustering is an un-supervised algorithm, that partitions the data-set into n singleton nodes and keeps merging mutually close pair of nodes until one final node is generated, resulting into a tree structure also referred as *Dendrogram*. We implemented the Seq2SeqAE and Seq2SeqVAE model with two different input transformations: one-hot encoding and word2vec [29] encoding. While HAC used GloVe embedding [30] for transforming the input data. For Seq2SeqAE and Seq2SeqVAE the encoder had an embedding layer and a mean layer for compressing the embedding output. Apart from two additional layer added to the encoder layer in case of word2vec the rest of the model was the same as for one-hot encoding. We trained the model with $bi, tri, 4, 5$-grams of feature sequences labelled using the proposed schema in this work for 100 epochs and the results are illustrated next. For reproduce-ability the code and the data used in this work can be accessed at the following link. [3]

IV. Results

The loss value of the baseline Seq2SeqAE with 5-grams of one-hot encoding was (0.2) and with word2Vec embedding was (4.0), while Seq2SeqVAE loss value was close to (0.08) and for word2vec embedding was (2.0) much less than the baseline model. Since, the loss values may tell very little about these generative models, clustering and analysing the latent space can reveal the difference much better.

To evaluate whether our labelling schema can assist the model in generating generic longer sequences we manually analysed the generated latent space and compare the Seq2SeqVAE with the Seq2SeqAE baseline model. The latent space was decomposed by independent component analysis (ICA) [31] in conjunction with principal component analysis (PCA) [32]. The results of decomposed latent space was compared with original data using K-means and Hierarchical Agglomerative Clustering Algorithms [33].

The clusters of original and decomposed latent space were then evaluated using Sklearn's [34] implementation of Silhouette [35] and Davis-Bouldin Index [36]. Silhouette score s is given by: $s = \frac{b-a}{max(a,b)}$ where, a is intra-cluster distance and b is inter-cluster distance. Davies Bouldin score db is given by: $db = \frac{1}{|C|}\sum_{i=1}^{C} \max_{i \neq j}(\frac{(Diam(C_i)+Diam(C_j))}{Dist(C_j,C_j)})$ where, i, j are clusters from the same partitioning, $|C|$ is the number of clusters, $Dist(C_i, C_j)$ is the inter-cluster distance, and $Diam(C_i)$ is the intra-cluster diameter. Table II indicates that Seq2SeqVAE out-performed the task of generating good clusters for all the sequences– *tri-/four-/five-* grams compared to Seq2SeqAE. The decomposition of features with PCA and ICA showed considerable reduction in the performance of the clustering results as highlighted in Table II fourth column, compared to non-decomposed input features in the third column.

Seq2SeqVAE model generates four clusters of the latent space using t-SNE with high inter-cluster distance in Sub-figure 3a with mixed data-points compared to the Seq2SeqAE baseline model that provides four big clusters and three small clusters with low inter-cluster distance(only the clustering results with good clusters are reported here). In Seq2SeqVAE it populates all the four clusters with medium to low sparsity. Task-driven human-human dialogue data-points(in green) have low sparsity and dispersed in all the clusters for Seq2SeqVAE. Task-driven human-machine and chitchat human-machine dialogues(blue and purple) are densely present close to each other in bottom and right cluster for Seq2SeqVAE. K-means clusters for 4-grams and tri-grams indicated in Sub-figures 3b, 3c presented the best outcome with respective KL-gains $(0.01, 0.1)$. A subset of data-points were selected from Seq2SeqVAE, Seq2SeqAE and HAC model's latent space. HAC model's latent space could provide sub-structures with around 5 segments that were meaningful, however their relation with the other sub-structures in its latent space could-not be established, hence failing in generalising long sequences. Seq2SeqAE provided more generalisation for sub-structures compared to HAC model in the sense that it could cluster sub-structures, however some clusters were very small and a manual analysis of them indicated high variability among the data-points, hinting towards lack of appropriate generalisation between long sequences. The Seq2SeqVAE presented superior clusters(in terms of generalising capability and the quality) based on the analysis of its latent space compared to the baseline models. Figure 4 presents a tree-like structure of dialogue features used as input to the Seq2SeqAE and Seq2SeqVAE, starting from right each segment of a dialogue utterance can be considered as a single input, which is labelled with dialogue control functions. These functions are then labelled with SE and DP labels. A final annotation of all the input is done with opening, on-going and closing labels. The generalisation capacity displayed in the latent space of Seq2SeqVAE, Seq2SeqAE, and HAC are represented with blue, white and grey colour boxes, respectively. These regions indicate the depth of the annotation schema that could be generated by the models and the Seq2SeqVAE generated sub-structures with greatest depth. Our annotation schema seems promising as the results indicate that the Seq2SeqVAE model was able to learn long sequences (hierarchical labels with $tri, four, five$-grams of segments) and was able to generalise them to form clusters.

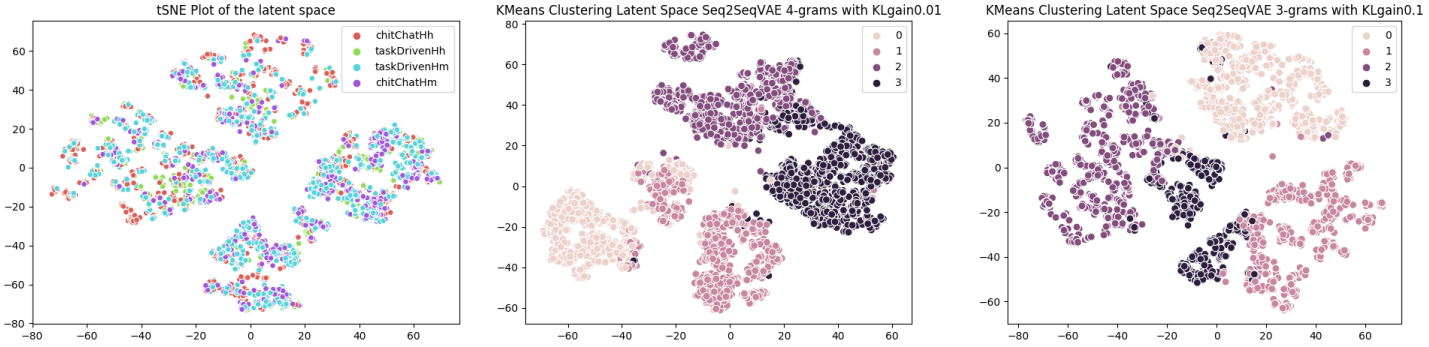
V. Summary, Discussion and Future Work

This work proposed a novel schema of sub-structuring using syntactic features, dialogue control functions, sequence expansion and dialogue policies. In order to find generic sub-structures based on our proposed schema we combine corpus of four different types from four sources. The features (selected labels or entire schema of labels) are transformed to one-hot encoding and word2Vec embedding and used an

[3] https://github.com/maitreyeeT/Seq2SeqDialStruct.git

Model/Features	KL-gain	Silhouette Score	Silhouette score (PCA+ICA)	Davies-Bouldin	Davis-Bouldin (PCA+ICA)
1.*VAE*/*tri*-gram	*1*	HAC=*0.54*, Kmeans=*0.52*	HAC=*0.1*, Kmeans=*0.1*	HAC=*0.59*, Kmeans=*0.56*	HAC=*3.2*, Kmeans=*2.2*
2.*VAE*/*tri*-gram	*0.1*	HAC=*0.45*, Kmeans=*0.46*	HAC=*0.15*, Kmeans=*0.12*	HAC=*0.71*, Kmeans=*0.63*	HAC=*2.3*, Kmeans=*2.2*
3.*VAE*/*tri*-gram	*0.01*	HAC=*0.60*, Kmeans=*0.54*	HAC=*0.09*, Kmeans=*0.12*	HAC=*0.54*, Kmeans=*0.55*	HAC=*3.2*, Kmeans=*2.3*
4.*AE*/*tri*-gram	–	HAC=*0.28*, Kmeans=*0.24*	HAC=*0.16*, Kmeans=*0.23*	HAC=*1.5*, Kmeans=*1.3*	HAC=*1.4*, Kmeans=*2.2*
5.*VAE*/4-gram	*1*	HAC=*0.6*, Kmeans=*0.48*	HAC=*0.13*, Kmeans=*0.13*	HAC=*0.58*, Kmeans=*0.61*	HAC=*1.9*, Kmeans=*2.3*
6.*VAE*/4-gram	*0.1*	HAC=*0.59*, Kmeans=*0.53*	HAC=*0.15*, Kmeans=*0.15*	HAC=*0.54*, Kmeans=*0.57*	HAC=*2.5*, Kmeans=*2.2*
7.*VAE*/4-gram	*0.01*	HAC=*0.63*, Kmeans=*0.49*	HAC=*0.09*, Kmeans=*0.13*	HAC=*0.58*, Kmeans=*0.59*	HAC=*3.4*, Kmeans=*2.3*
8.*AE*/4-gram	–	HAC=*0.31*, Kmeans=*0.22*	HAC=*0.24*, Kmeans=*0.18*	HAC=*1.4*, Kmeans=*1.2*	HAC=*1.8*, Kmeans=*1.6*
9.*VAE*/5-gram	*1*	HAC=*0.57*, Kmeans=*0.49*	HAC=*0.16*, Kmeans=*0.12*	HAC=*0.56*, Kmeans=*0.59*	HAC=*1.3*, Kmeans=*2.3*
10.*VAE*/5-gram	*0.1*	HAC=*0.52*, Kmeans=*0.48*	HAC=*0.17*, Kmeans=*0.14*	HAC=*0.63*, Kmeans=*0.58*	HAC=*1.4*, Kmeans=*2.2*
11.*VAE*/5-gram	*0.01*	HAC=*0.56*, Kmeans=*0.51*	HAC=*0.12*, Kmeans=*0.12*	HAC=*0.61*, Kmeans=*0.57*	HAC=*2.3*, Kmeans=*2.2*
12.*AE*/5-gram	–	HAC=*0.32*, Kmeans=*0.23*	HAC=*0.24*, Kmeans=*0.19*	HAC=*1.30*, Kmeans=*1.29*	HAC=*1.8*, Kmeans=*1.5*

TABLE II: The first column in this Table gives the information about the model and the features used, second column presents the hyper-parameter used, while the rest of the columns provides evaluation scores for hierarchical (HAC) and K-means clustering algorithms' performed on the original latent space (columns 3 and 5) and decomposed latent space (columns 4 and 6)



(a) t-SNE plot of Seq2SeqVAE for 5-grams (b) t-SNE plot of Seq2SeqVAE for 4-grams (c) t-SNE of Seq2SeqVAE for 3-grams

Fig. 3: Clustering of the Seq2SeqVAE and Seq2SeqAE baseline model's latent space after 1000 epocs.

input to our Seq2SeqVAE. The results show that Seq2SeqVAE model out-performed the baselines Seq2SeqAE and hierarchical agglomerative clustering.

Enabling human like dialogues with machines is non-trivial and most of the practical models in the research community have explored limited scenarios with specific or no structuring features. Contrary to the most common practices we believe structuring of dialogue features based on the latent space of generative models such as autoencoders can allow machines to manage and generate diverse dialogue functions in-order to step towards mixed-initiative and human-like dialogues. The structure labelling schema provided is promising in order to have longer sub-structures in the latent space.

One of the limitation of this work is derived relation and classes of clusters could be biased, as the latent space was manually analysed and there is sparsity and ambiguity surrounding the standardised evaluation metrics for unsupervised methods used in this work. Other limitation of this work is with-respect to imbalance and small set of dialogue samples for different genres, and for a more robust model that can generalise requires samples from other genres and also a much larger corpus. Since, the focus of this work was to evaluate the model for long sub-structures in dialogues rather than the annotation schema, hence a final issue could raise questions on its validity and necessity.

Possible solutions that can be employed to mitigate the limitations of the model in the future is to primarily increase the diversity and the quantity of the corpus. Secondly, data augmentation (up and down sampling) and knowledge creation (co-referencing, pronoun resolution, and context modelling by improving the attention layer using syntactic or semantic dependency graphs) can be used for creating semantic sequences instead of using raw syntactic features from dependency graphs and an attention layer based on previous input. In order to evaluate the annotation schema, the results of the model can be either compared for the input with annotations to that without or with another generic annotation schema. An important immediate future work is to analyse the relation between clusters of the latent space and dialogue flows in the corpus. In the near future this work will be used to build DMS strategies either through a hand-crafted method or with Bayesian networks based on the generated latent space.

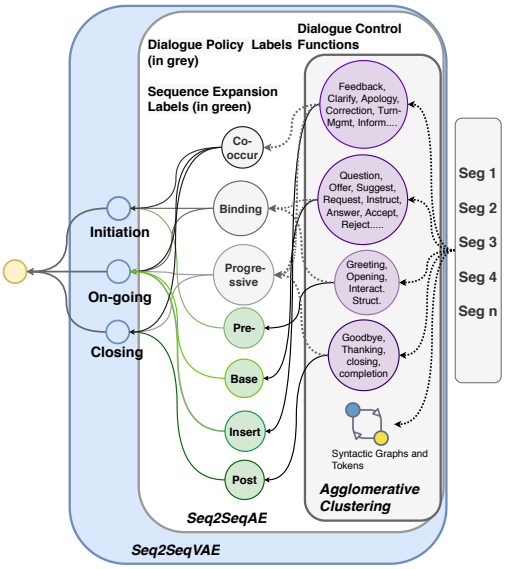


Fig. 4: Generalisation capability of all the tested models for the presented annotation schema is shown here.

ACKNOWLEDGMENTS

We would like to thank Prof. Thomas Hellström and Associate Prof. Suna Bensch from the Department of Computing Science at Umeå University for their valuable insights during some of the research discussions.

REFERENCES

[1] M. B. Hoy, "Alexa, siri, cortana, and more: an introduction to voice assistants," *Medical reference services quarterly*, vol. 37, no. 1, pp. 81–88, 2018.

[2] A. S. Tulshan and S. N. Dhage, "Survey on virtual assistant: Google assistant, siri, cortana, alexa," in *International Symposium on Signal Processing and Intelligent Recognition Systems*. Springer, 2018, pp. 190–201.

[3] J. Alexandersson and N. Reithinger, "Learning dialogue structures from a corpus," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[4] S. Bangalore, G. Di Fabbrizio, and A. Stent, "Learning the structure of task-driven human–human dialogs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 7, pp. 1249–1259, 2008.

[5] R. C. Gunasekara, D. Nahamoo, L. C. Polymenakos, J. Ganhotra, and K. P. Fadnis, "Quantized-dialog language model for goal-oriented conversational systems," *arXiv preprint arXiv:1812.10356*, 2018.

[6] Z. Ales, A. Pauchet, and A. Knippel, "Extraction and clustering of two-dimensional dialogue patterns," *International Journal on Artificial Intelligence Tools*, vol. 27, no. 02, p. 1850001, 2018.

[7] T. D. Midgley, S. Harrison, and C. MacNish, "Empirical verification of adjacency pairs using dialogue segmentation," in *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*. Sydney, Australia: Association for Computational Linguistics, 2006, pp. 104–108.

[8] N. Duran and S. Battle, "Conversation analysis structured dialogue for multi-domain dialogue management," in *DEXAHAI*, 12 2018, p. 4.

[9] T. Bub and J. Schwinn, "Verbmobil: The evolution of a complex large speech-to-speech translation system," in *In Int. Conf. on Spoken Language Processing*, 1996, pp. 2371–2374.

[10] M. Frampton and O. Lemon, "Reinforcement learning of dialogue strategies using the user's last dialogue act," in *IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, 2005.

[11] S. Emanuel A and S. Harvey, "Opening up closings," *Semiotica*, vol. 8, no. 4, pp. 289–327, 1973.

[12] W. Shi, T. Zhao, and Z. Yu, "Unsupervised dialog structure learning," 2019.

[13] M. Tewari and S. Bensch, "Natural language communication with social robots for assisted living," in *IROS Workshop in Robots for Assisted Living*, Madrid, Spain, 2018, p. 4.

[14] M. Tewari, "Beyond adjacency pairs: Hierarchical clustering of long sequences for human-machine dialogues," in *1st Workshop on Computational Approaches to Discourse 2020- EMNLP Workshop (Accepted)*, 2020, pp. 1–9.

[15] J. Sidnell and T. Stivers, *The handbook of conversation analysis*. John Wiley & Sons, 2012, vol. 121.

[16] S. Tanya, *Sequence Organization*. UK: Wiley-Blackwell, 2012, ch. 10, pp. 191–209.

[17] H. Bunt, V. Petukhova, A. Malchanau, A. Fang, and K. Wijnhoven, "The dialogbank: dialogues with interoperable annotations," *Language Resources and Evaluation*, vol. 53, no. 2, pp. 213–249, 2019.

[18] H. Ryuichiro, K. Funakoshi, M. Inaba, Y. Tsunomori, T. Takahashi, and K. Nobuhiro, "Overview of dialogue breakdown detection challenge 3," *Proceedings of Dialogue System Technology Challenge*, p. 14, 2017.

[19] H. S. Thompson, A. Anderson, E. G. Bard, G. Doherty-Sneddon, A. Newlands, and C. Sotillo, "The hcrc map task corpus: natural dialogue for speech recognition," in *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1993, pp. 25–30.

[20] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ser. ICASSP'92. USA: IEEE Computer Society, 1992, p. 517–520.

[21] J. F. Allen, L. K. Schubert, G. Ferguson, P. Heeman, C. H. Hwang, T. Kato, M. Light, N. Martin, B. Miller, M. Poesio *et al.*, "The trains project: A case study in building a conversational planning agent," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 7, no. 1, pp. 7–48, 1995.

[22] L. Per, *Troubles with Mutualities: Towards a Dialogical Theory of Misunderstanding and Miscommunication*. Cambridge University Press, 1995, ch. 8, pp. 176–212.

[23] H. Bunt, "Dimensions in dialogue act annotation," in *LREC*, 2006.

[24] S. Gupta, R. Shah, M. Mohit, A. Kumar, and M. Lewis, "Semantic parsing for task oriented dialog using hierarchical representations," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Belgium: ACL, 2018, p. 6.

[25] M. Tewari, "Formalization of dialogues with cooperating distributed grammar systems," Umeå University, Department of Computing Science, Tech. Rep. 2020:9, 2020.

[26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Generative Models; Variational Autoencoder*. MIT press, 2016, ch. 20.

[27] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proceedings of ICML workshop on unsupervised and transfer learning*, 2012, pp. 37–49.

[28] I. Goodfellow, Y. Bengio, and A. Courville, *Sequence modeling: recurrent and recursive nets*. MIT press, 2016, ch. 10.

[29] M. Tomas, S. Ilya, C. Kai, C. Greg S, and D. Jeffrey, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'13. USA: Curran Associates Inc., 2013, pp. 3111–3119.

[30] P. Jeffrey, S. Richard, and M. Christopher D, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: ACL, 2014, pp. 1532–1543.

[31] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[32] J. Shlens, "A tutorial on principal component analysis," *arXiv preprint arXiv:1404.1100*, p. 12, 2014.

[33] X. Dongkuan and T. Yingjie, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, Jun 2015.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[35] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53 – 65, 1987. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0377042787901257

[36] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.