Unsupervised Flow Discovery from Task-oriented Dialogues

Patrícia Ferreira[1,2], Daniel Martins[3], Ana Alves[1,3], Catarina Silva[1,2], and Hugo Gonçalo Oliveira[1,2]

[1] CISUC, Universidade de Coimbra, Portugal
[2] DEI, FCTUC, Universidade de Coimbra, Portugal
[3] ISEC, Instituto Politécnico de Coimbra, Portugal

Abstract. The design of dialogue flows is a critical but time-consuming task when developing task-oriented dialogue (TOD) systems. We propose an approach for the unsupervised discovery of flows from dialogue history, thus making the process applicable to any domain for which such an history is available. Briefly, utterances are represented in a vector space and clustered according to their semantic similarity. Clusters, which can be seen as dialogue states, are then used as the vertices of a transition graph for representing the flows visually. We present concrete examples of flows, discovered from MultiWOZ, a public TOD dataset. We further elaborate on their significance and relevance for the underlying conversations and introduce an automatic validation metric for their assessment. Experimental results demonstrate the potential of the proposed approach for extracting meaningful flows from task-oriented conversations.

Keywords: Natural Language Processing · Task-oriented Dialogues · Dialogue Flow · Unsupervised Flow Discovery · Dialogue Guidance

1 Introduction

Dialogue systems are ubiquitous today. Customer-support agents are used by companies to handle customer service inquiries and provide assistance; virtual assistants like Siri[4] and Alexa[5] are integrated into a range of smart devices to provide daily assistance through natural language interactions.

The aforementioned conversational agents are known as task-oriented systems, designed to fulfill distinct user tasks or requirements, such as making a reservation or requesting information. To cover all the target tasks, along with all possible interactions, creating dialogue flows for these systems is complex, time-consuming and often done manually. Furthermore, the resulting flows can rarely be shared across domains, meaning that their design has to be performed for each new dialogue system.

[4] https://www.apple.com/siri/

[5] https://alexa.amazon.com

This is where the automatic discovery of dialogue flows can play a key role. By analysing the history of conversational turns, this task may help in the identification of communication trends, which can support the design of flows for virtual agents, thus accelerating their processing, or help with their validation, possibly suggesting changes. Moreover, such flows enable to track the progression of the conversation over time and guide human agents, e.g., in a call centre, helping them to anticipate questions and provide prompt responses, without losing the benefits of human interaction [13].

We propose an approach for automatically discovering dialogue flows from a history of conversations. It is unsupervised, thus applicable to any collection of task-oriented dialogues (TODs), and follows three main steps. Utterances are (i) represented in a vector space and (ii) clustered according to their semantic similarity; (iii) discovered clusters, which may be seen as dialogue states, are used as vertices of a transition graph. For human consumption, the graph can be visually depicted, incorporating computed transitions and their corresponding probabilities, derived from the conversation history. For easier interpretation, states are labelled based on the most frequently occurring sequences within the clustered utterances. This visual representation empowers users to conduct a post-inspection of the discovered flow, also useful for auditioning.

Towards its validation, the proposed approach is implemented and applied to a well-known dataset of TOD, MultiWOZ [3]. Resulting dialogue flows are depicted, which allows for a comprehensive understanding of the discovered patterns. Furthermore, we propose an automatic measure for computing to what extent the transitions in the test portion of the dataset align with the flows discovered from the training portion. This establishes the reliability and accuracy of our approach.

The remainder of the paper is structured as follows: Section 2 reviews previous work related to dialogue flow discovery; Section 3 describes the proposed approach; Section 4 describes how it was experimented, including the used dataset and implementation details; Section 5 presents the resulting flows, followed by an attempt to their automatic evaluation; Section 6 concludes the paper and provides cues for future work.

2 Related Work

Utterances in dialogues are commonly classified according to: (i) user intents, e.g., make a reservation, find attractions; or (ii) dialogue acts, more generic, relative to the action performed by the speaker, e.g., ask, explain, request.

To some extent, the significance of the previous labels can be overlooked by data-driven chatbots, such as sequence-to-sequence agents [17], also including agents built on top of large language models, e.g., ChatGPT[6] or Bard[7]. However, such labels are highly relevant for task-oriented systems, which rely on the

[6] `https://chat.openai.com/`

[7] `https://bard.google.com/`

dialogue structure to achieve specific goals efficiently. Their incorporation facilitates successful interactions and goal completion, setting them apart from more general conversational models.

The automatic classification of utterances is typically done in a supervised fashion [4,5], thus requiring annotated data, which may not be readily accessible in numerous domains. Hence, the development of an approach that does not rely on annotated data has the potential to significantly expand the range of application scenarios.

The design of dialogue flows in task-oriented systems can steer the conversation in specific directions, avoiding purely reactive responses [6]. It encompasses the definition of task-specific intents and of training phrases, among other decisions, and generally ends up being created manually, often with the help of tools like Google's DialogFlow[8], Microsoft Luis[9], or Rasa[10].

To automate this process, it is necessary to cluster expressions that represent the same intent, the same action, or mention the same information. For this, utterances are typically represented in a vector space where those semantically similar are closely positioned together. Methods like word [8] or sentence [11,12] embedding serve this purpose, enabling to group utterances with comparable meanings or content.

Towards intent discovery, utterances may be clustered with well-established algorithms like k-means, directly applied to the utterance embeddings [12,11], possibly incorporating additional contextual features [19].

Some related works represent dialogue flows as graphs, with the discovered clusters as the vertices [9,2,14]. An earlier approach [14] tackles the same problem, employing a Hidden Markov Model (HMM) when learning from Twitter conversations. Their approach introduces interesting features, such as the vertices for representing the start and end of dialogue, and a threshold for ignoring transitions with low probability. Clusters were labelled manually, while our work aims to streamline and enhance the efficiency of dialogue flow discovery, also reducing the manual efforts for cluster labelling.

Clusters group similar utterances without considering the sequence of conversation [9]. But graphs may represent the main transitions between topics and dialogues [2]. Different methods were proposed to model such transitions, e.g., HMM [14], Latent Dirichlet Allocation [10] or Deep Learning techniques [16].

Graph2Bots [1] is a tool for assisting agent conversation designers in extracting graphs of human-to-human conversations. The resulting graph includes the main dialogue phases and their transitions. An alternative approach finds the most frequent sequences of questions and responses when representing their sequences by a finite-state automaton [15]. Despite relying on intent labels and on proprietary software, the previous work enables to visualize the flows and was demonstrated for the restaurant booking dialogues in the MultiWOZ dataset. None of the previous undergoes a quantitative evaluation of the resulting flows.

[8] https://cloud.google.com/dialogflow/

[9] https://www.luis.ai/

[10] https://rasa.com/

We propose both a new approach for flow discovery and a metric for flow validation in unseen dialogues. Given a history of dialogues, similar utterances are clustered; a graph is created based on cluster transitions; and a visualisation is provided, to support TOD systems and their users.

3 Unsupervised Dialogue Flow Discovery

We propose a generic approach for automatically discovering the most frequent flows in a history of dialogues, and for representing them visually, as a transition graph. Figure 1 summarises this approach, which encompasses three steps: utterance representation, utterance clustering, and flow discovery.

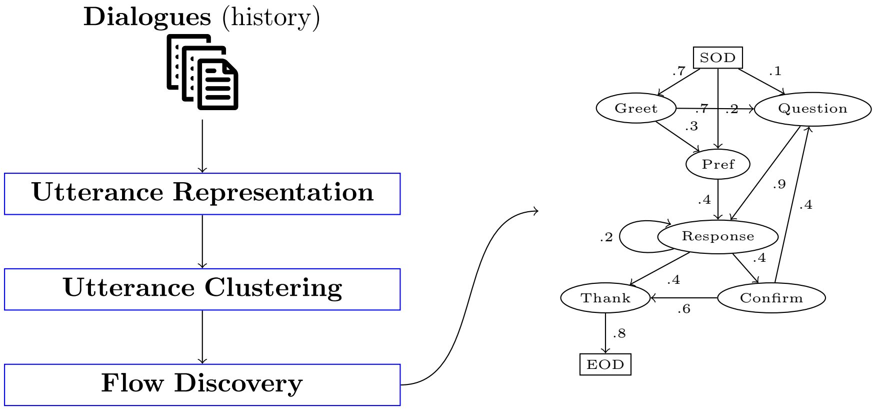

Fig. 1. Overview of the proposed approach with three steps: utterance representation, utterance clustering, and flow discovery.

In order to apply our approach to a specific domain, a history of dialogues covering all the relevant intents for that domain is required. Dialogues are represented as written sequences of utterances contributed by various participants. While our approach is versatile and can be adapted to various scenarios, our primary focus is on TOD involving a customer and an agent.

In Figure 1, from top to bottom, in the first step, the dialogues are represented according to their meaning, for example, through their embeddings. In the second step, the dialogues are grouped based on their semantic similarity, with the intention that these groupings represent dialogue states. In the third step, transitions between the states and their corresponding probabilities are calculated. This results in a transition graph $G(C, T)$, with the states $c \in C$ as vertices and transitions $t(c_a, c_b, p_{ab}) \in T$ as edges, where p_{ab} is the probability of transitioning from state c_a to state c_b. The latter is calculated according to Equation (1), where $|t(c_a, c_b)|$ represents the number of dialogues in c_b that immediately follow a dialogue in c_a.

$$p_{ab} = \frac{\|t(c_a, c_b)\|}{\sum_{x \in C} \|t(c_a, c_x\|} \tag{1}$$

In addition to the vertices derived from the clusters, two vertices representing the Start of Dialogue (`SOD`) and the End of Dialogue (`EOD`) are included. Moreover, to ease human interpretation, states can be identified by automatically-assigned human-labels. These can be based on frequent sequences or keywords used in the clustered utterances.

4 Experimentation Setup

This section describes the experimentation designed for testing the proposed approach, focused on its application to a dataset of TOD. After described this dataset, implementation details are provided.

4.1 Dataset

Experimentation was conducted with MultiWOZ 2.2 [18], a well-known dataset of TOD. Its dialogues include multiple turns of interactions between two humans: a USER plays the role of a user, who has a task to fulfill; a SYSTEM tries to answer requests promptly, while helping the user to conclude the task. Dialogues cover a total of eight domains (restaurant, attraction, hotel, taxi, train, bus, hospital, police) and several intents, such as finding a restaurant or booking a taxi.

Besides the domain, other annotations (e.g., intent, slots) are available, but none of them are used in this work. This kind of information is rarely available, and would have to be predicted by an automatic classifier. Moreover, since intents change according to the domains of the dialogue, considering them would make our approach no longer domain-agnostic.

MultiWOZ 2.2 is divided into train and test portions, distributed according to Table 1. Dialogues are sequences of turns, each including a USER utterance and a SYSTEM response. In this work, flows are discovered from the train portion, with the test portion held out for validation. Table 2 illustrates the contents of the dataset with a single dialogue.

Table 1. Contents of MultiWOZ 2.2 dataset.

Portion	#Dialogues	#Utterances		
		Total	USER	SYSTEM
Train	8,436	113,552	56,776	56,776
Test	1,000	14,744	7,732	7,732

Table 2. Dialogue SNG1066 from the MultiWOZ 2.2 test portion.

Speaker	**Utterance**
USER	Could you help me find a boat to visit on the north side?
SYSTEM	I have one in that area. It's called the Riverboat Georgina. It's located at Cambridge Passenger Cruisers, Jubilee House. Would you like their phone number for more information?
USER	Yes, I want the phone number and also the entrance fee, please.
SYSTEM	Their phone number is 01223902091 and we do not have the entrance fee in our database at this time.
USER	Okay, that is all I need today. Thank you very much.
SYSTEM	You're very welcome! Thanks for contacting the Cambridge TownInfo Centre and have a great day!

4.2 Implementation

Different methods were tested for utterance representation, clustering and labelling. Yet, except for the latter, we focus on a single implementation and leave the comparison of alternatives for future work.

As long as it allows to measure the semantic similarity between different utterances, further enabling similarity-based clustering, any vector representation can be adopted for the utterances. We used the 384-sized embeddings obtained directly from the model `all-MiniLM-L6-v2`, available from the Sentence Transformers library[11].This is reported as one of the best-performing models for sentence embedding and also one of the smallest (80MB).

Any clustering algorithm could have been used in the second step, but we chose the k-means algorithm, available in the scikit-learn library[12], due to its widespread use in clustering tasks. k-means was used with a maximum of 1,000 iterations and centroids initialized by sampling, based on an empirical probability distribution of points (`k-means++`). For MultiWOZ 2.2, the number of clusters, denoted as k, was determined with Silhouette method, which assesses the cohesion and separation of clusters.

As we always know who the speaker of each utterance is, user and system utterances were analyzed separately, resulting in a different number of clusters for each. A scatter plot of Silhouette scores for various values of k is illustrated in Figure 2. The maximum score is when $k = 18$ and $k = 13$, respectively for the user and the system. These were the numbers of clusters used.

On the flow discovery, besides computing the transitions and their probabilities as in Section 3, two approaches were tested for labelling the clusters. One uses the most frequent verb phrases (VPs) in their utterances, obtained from the model `en_core_web_md` of the spaCy[13] toolkit; the other relies on KeyBERT [7] for extracting the most frequent keywords in the utterances of the cluster.

[11] https://www.sbert.net/

[12] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[13] https://spacy.io/

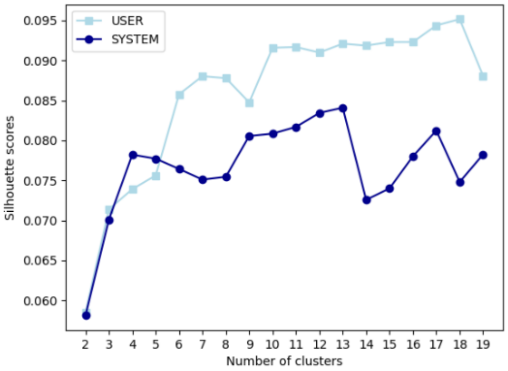

Fig. 2. Silhouette scores for different values of k for USER and SYSTEM utterances.

5 Results

This section summarizes the results obtained with the previously described implementation of the approach. It presents resulting dialogue flows and describes an attempt at the automatic evaluation of their effectiveness.

5.1 Discovered Flows

Following the procedure described in Section 4.2, utterance clustering resulted in 18 USER and 13 SYSTEM clusters (i.e., states). These could be identified by a number and simply listed, but this would not provide enough information on the dialogue state.

However, dialogue states, alone, are not enough for representing the flows, which should include transitions between states. As mentioned in Section 3, dialogue flows can be represented by graphs, but the ideal visualisation is not straightforward. Besides the challenge of generating informative labels, a graph with all the states and transitions will contain some rare transitions and can be too complex to manage. On the other hand, a graph with few states will be easier to read at the cost of losing coverage of a fraction of transitions. Therefore, we look at graphs after the application of different thresholds θ on the transitions.

Figures 3 and 4 depict the discovered flows, respectively for $\theta = 0.10$ and $\theta = 0.15$. Vertices in light blue represent the USER; in dark blue represent the SYSTEM; and in yellow correspond to the start (`SOD`) and end (`EOD`) of dialogue.

Labels give insight on the system's actions and responses, such as hotel suggestions, available options or greetings, and reveal user's needs and intentions, such as hotel search, help and specific information. If different clusters end up with the same label, a unique number is appended. The VP method tends to result in grammatically well-constructed labels, whereas KeyBERT relies on the most relevant terms.

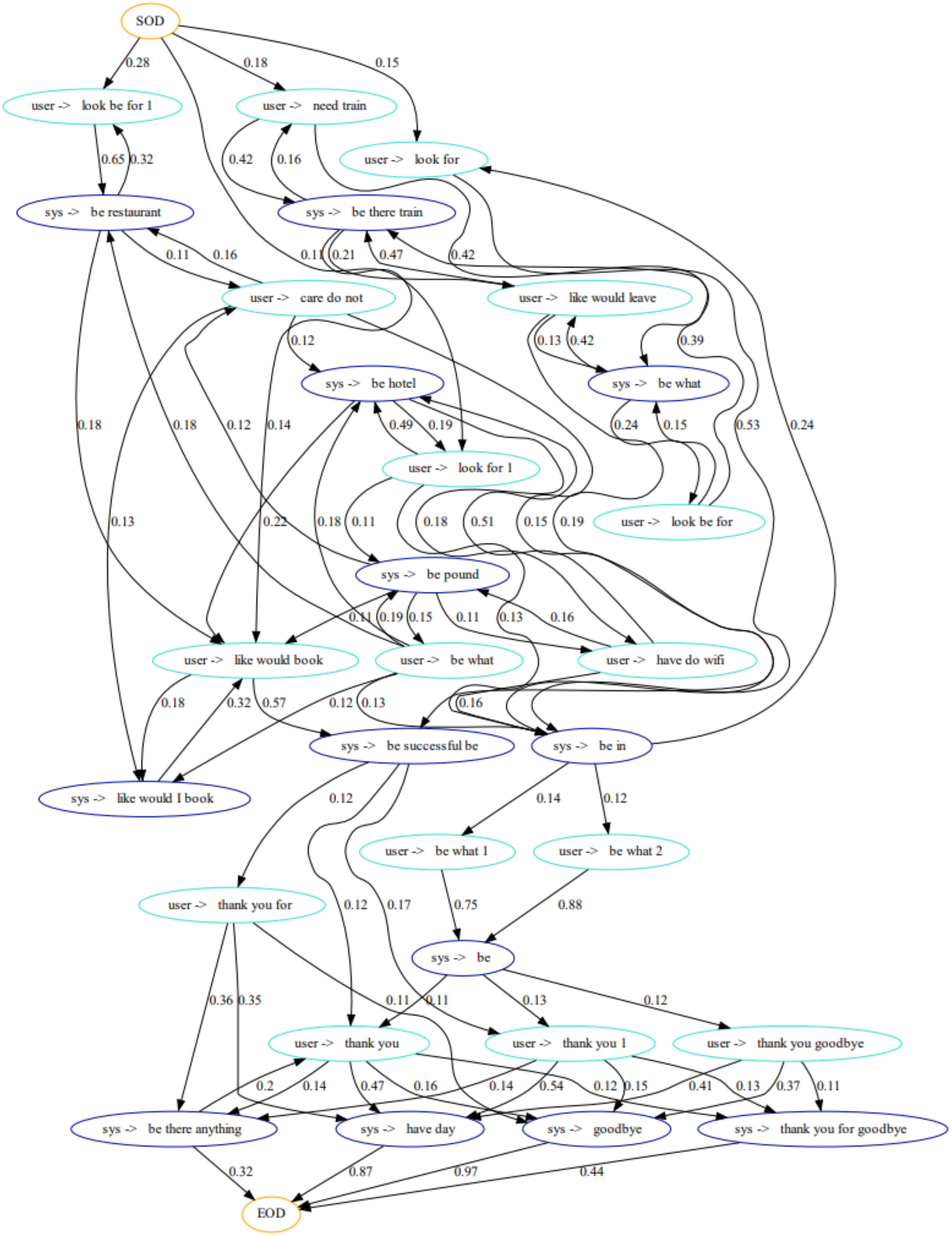

Fig. 3. Flow discovered from the MultiWOZ train portion with $\theta = 0.10$ and labels generated from the most frequent verb phrases.

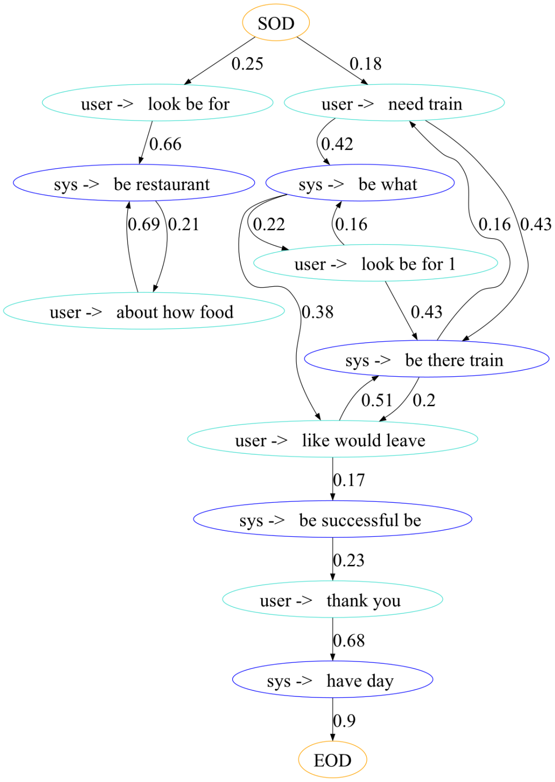

Fig. 4. Flow discovered from the MultiWOZ train portion with $\theta = 0.15$ and labels generated from the most frequent verb phrases.

By increasing the value of θ, lower-probability transitions are ignored, resulting in more simplified graphs focused on the strongest relationships between clusters. Conversely, by reducing the value of θ, less frequent transitions can be included in the visualisation, providing more detailed insights into the possible interactions that may occur during the dialogue.

When analysing these flows, we notice common patterns in the dialogues of MultiWOZ. They usually begin with a request or a question by the user, related to their specific need, but generally without an isolated greeting, common in social interactions. This might be due to the task-oriented nature of the dataset, where users have clear goals and seek prompt answers to their needs.

As the dialogue progresses, the interaction between the user and the system unfolds, with exchanges of information, answers and clarifications to meet the user's needs. This intermediate phase may include several steps, such as the system providing options or additional details, and the user making choices or asking for more specific information.

Finally, upon reaching the end of the dialogue, it is common to find a conclusion or fulfilment step, where the system may thank the user for the interaction, provide final information, or say goodbye.

5.2 Automatic Evaluation

Flow visualizations provide interesting insights, but drawing conclusions on their utility and on how well they represent the underlying dialogues is challenging, especially without experts in the dataset's domains and call-center procedures. Yet, quantitatively evaluating flows is complex and lacking in related work [1,15].

Towards a quantitative metric of success, we propose to assess the flow by measuring to what extent an unseen portion of dialogues (i.e., test set) follows the flows discovered from another portion of the same type (i.e., train set). Briefly, this evaluation encompasses the following steps:

1. Discover flows from dialogues in the train set.
2. For each utterance u_x in the test set:
   (a) assign u_x to the most similar cluster in the flow, $c_x \in C$;
   (b) get the utterance that directly follows u_x, u_y;
   (c) assign u_y to the most similar cluster in the flow, $c_y \in C$;
   (d) check whether the transition $t(c_x, c_y, p_{xy})$ is in the flow (i.e, is predicted).

For different values of θ, we calculated the **transition accuracy**, i.e., the proportion of predicted transitions in relation to all the transitions in the test set. Obtained values are in Table 3.

The higher θ is, the smaller the graph, and the lower the accuracy. With $\theta = 0.20$, there are no paths between `SOD` and `EOD`, so no transitions are predicted. With $\theta = 0.05$, more than 80% of the transitions in the test portion are predicted. This is promising, especially considering that it is based on unseen dialogues, with previously unseen utterances, where language may vary. Nevertheless, we should note that requests and responses in MultiWOZ tend to be direct. And still, we see a significant number of transitions ($\approx$ 18%) that follow rare paths.

Table 3. Configuration of the dialogue flows discovered from the MultiWOZ 2.2 train portion, with different values of threshold θ; and transitions accuracy in the MultiWOZ 2.2 test portion, as computed by the previous flow.

θ	$\lvert C\rvert$	$\lvert T\rvert$	**Accuracy**
0.05	33	139	82.2%
0.10	32	66	70.4%
0.15	23	32	43.6%
0.20	0	0	0

This score is not comparable to an evaluation by human experts, but it is a fast way of obtaining initial quality-related figures. Moreover, it enables the comparison of flows discovered with different parameters and approaches. In the future, we will attempt to increase accuracy, by testing different embeddings or clustering algorithms, ideally without increasing the number of clusters.

6 Conclusion and Future Work

Motivated by the importance of flows in TOD systems, and by the amount of manual work involved in their production, validation, and maintenance, we presented an innovative approach for unsupervised flow discovery. To provide a comprehensive analysis of communication trends, it follows a three-step method that: represents utterances as vectors; clusters semantically similar ones to form dialogue states; and constructs a transition graph with such states as vertices.

Experiments showcase the potential of our approach, providing a pathway for extracting meaningful dialogue flows without the need for annotated data. The graphical visualisations offer an intuitive representation, supporting both developers and users in interpreting and optimising TOD systems. This is complemented with the proposal of an automatic evaluation, which shows that the flows represented by 33 dialogue states can predict more than 80% of the transitions in the test portion of MultiWOZ 2.2, a widely-used dataset in the development of TOD systems.

There is room for improvement, but progress can now be measured with the proposed metric. For clustering, it would be interesting to: explore other algorithms, including density-based, which do not require the number of clusters as input; consider the context of previous utterances as contextual features; extensively assess the impact of different models for utterance embedding. Another line would be investigating different methods for labelling clusters and visualising dialogue flows, towards deeper insights and more informative representations. The approach should be further validated in diverse data, including less artificial dialogues. We are currently working with industry partners and have plans to soon apply it to real call-center data. This should provide a more comprehensive evaluation of its effectiveness, and may be further complemented by a human-expert evaluation, hopefully offering valuable assessments and insights.

Towards more trustworthy machines, it would also be interesting to discover flows from virtual agents that are not flow-based, such as data-driven chatbots (e.g., ChatGPT and Bard), thus contributing to their explanation.

Acknowledgements This work was partially supported by: the project FLOWANCE (POCI-01-0247-FEDER-047022), co-financed by the European Regional Development Fund, through PT2020, and by the Competitiveness and Internationalization Operational Programme; the Portuguese Recovery and Resilience Plan through project C645008882-00000055, Center for Responsible AI; and by national funds through FCT, within the scope of the project CISUC (UID/CEC/00326/2020).

References

1. Bouraoui, J.L., Le Meitour, S., Carbou, R., Barahona, L.M.R., Lemaire, V.: Graph2bots, unsupervised assistance for designing chatbots. In: Procs. 20th Annual SIGdial Meeting on Discourse and Dialogue. pp. 114–117. ACL (2019)

2. Bouraoui, J.L., Lemaire, V.: Cluster-based graphs for conceiving dialog systems. In: Procs ECML-PKDD 2017 Workshop on Interactions between Data Mining and Natural Language Processing. CEUR-WS.org (2017)
3. Budzianowski, P., Wen, T.H., Tseng, B.H., Casanueva, I., Ultes, S., Ramadan, O., Gašić, M.: MultiWoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In: Procs. 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018. pp. 5016–5026 (2018)
4. Cohen, W., Carvalho, V., Mitchell, T.: Learning to classify email into "speech acts". In: Procs. 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 309–316 (2004)
5. Firdaus, M., Golchha, H., Ekbal, A., Bhattacharyya, P.: A deep multi-task model for dialogue act classification, intent detection and slot filling. Cognitive Computation **13**, 626–645 (2021)
6. Grassi, L., Recchiuto, C.T., Sgorbissa, A.: Knowledge-grounded dialogue flow management for social robots and conversational agents. International Journal of Social Robotics **14**(5), 1273–1293 (2022)
7. Grootendorst, M.: KeyBERT: Minimal keyword extraction with BERT. (2020). `https://doi.org/10.5281/zenodo.4461265`
8. Hashemi, H.B., Asiaee, A., Kraft, R.: Query intent detection using convolutional neural networks. In: International conference on web search and data mining, workshop on query understanding (2016)
9. Joty, S.R., Carenini, G., Lin, C.Y.: Unsupervised modeling of dialog acts in asynchronous conversations. In: Procs. 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011) (2011)
10. Li, X., Ouyang, J., Zhou, X., Lu, Y., Liu, Y.: Supervised labeled latent dirichlet allocation for document categorization. Applied Intelligence **42**, 581–593 (2015)
11. Liu, P., Ning, Y., Wu, K.K., Li, K., Meng, H.: Open intent discovery through unsupervised semantic clustering and dependency parsing. arXiv preprint arXiv:2104.12114 (2021)
12. Park, J., Jang, Y., Lee, C., Lim, H.: Analysis of utterance embeddings and clustering methods related to intent induction for task-oriented dialogue. arXiv preprint arXiv:2212.02021 (2022)
13. Rapp, A., Curti, L., Boldi, A.: The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. International Journal of Human-Computer Studies **151**, 102630 (2021)
14. Ritter, A., Cherry, C., Dolan, B.: Unsupervised modeling of twitter conversations. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. p. 172–180. HLT '10, ACL, USA (2010)
15. Sastre Martinez, J.M., Nugent, A.: Inferring ranked dialog flows from human-to-human conversations. In: Procs. 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 312–324. ACL, Edinburgh, UK (Sep 2022)
16. Serradilla, O., Zugasti, E., Rodriguez, J., Zurutuza, U.: Deep learning models for predictive maintenance: a survey, comparison, challenges and prospects. Applied Intelligence **52**(10), 10934–10964 (2022)
17. Vinyals, O., Le, Q.V.: A neural conversational model. In: Proceedings of ICML 2015 Deep Learning Workshop. Lille, France (2015)
18. Zang, X., Rastogi, A., Sunkara, S., Gupta, R., Zhang, J., Chen, J.: MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In: Procs. 2nd Workshop on Natural Language Processing for Conversational AI. pp. 109–117. ACL, Online (Jul 2020)
19. Zhang, H., Xu, H., Lin, T.E., Lyu, R.: Discovering new intents with deep aligned clustering. In: Procs. AAAI Conference on Artificial Intelligence. vol. 35 (16), pp. 14365–14373 (2021)