TaskSim: A Similarity Metric for Task-Oriented Conversations

Anonymous ACL submission

Abstract

The popularity of conversational digital assistants has resulted in efforts to improve user experience by extracting insights from the logs. These approaches utilize distance based metrics to identify similarities between user conversations. These metrics are typically designed to compare text snippets and do not take advantage of the unique conversational features in dialogues that are absent from other textual sources. To address this gap, in this work, we present *TaskSim*, a novel conversational similarity metric that utilizes different dialogue components (e.g.utterances, intents, and slots) with optimal transport. Extensive experimental evaluation of the *TaskSim* metric on a benchmark dataset demonstrate its superior performance over other traditional similarity approaches.

1 Introduction

Task-oriented conversational assistants have become increasingly popular in multiple industries enabling users to perform tasks such as travel reservations, banking transactions, online shopping, etc., through multi-turn conversations. These assistants support pre-defined sets of tasks that are executed based on user objectives or *intents*, and their associated task parameters or *slots* provided in the conversation (c.f. Figure 1).

The increased use of these assistants has led to the availability of valuable user-assistant conversation logs (Budzianowski et al., 2018; Andreas et al., 2020), prompting efforts to extract insights to improve the user-experience including personalized response generation, next-action recommendations, and information retrieval (Yaeli et al., 2022; Gao et al., 2020; Li et al., 2022). A key aspect of such conversational analytics is identifying similarities and dissimilarities between conversations.

Measuring semantic textual similarity is the basis of many natural language and text processing tasks, such as question answering, sentiment analysis, and information extraction (Zhou et al., 2015; Ye et al., 2016; Poria et al., 2016); it has been extensively studied for textual sources like documents, social media, transcripts, etc. However, there has been limited prior work studying similarity in task-oriented conversation settings (Appel et al., 2018; Lavi et al., 2021). Most approaches leverage popular word embeddings like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) to obtain higher-dimension semantic vector representations of words, and then use distance-based approaches such as cosine and edit-distance to compute the similarity between text snippets.

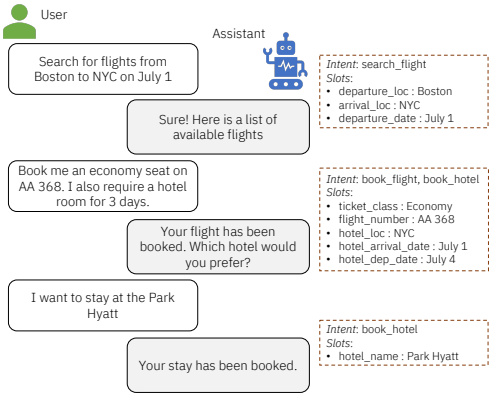

Figure 1: Sample task-oriented conversation depicting user intents, slot parameters, and their respective values.

While such approaches can identify semantic relationships between texts, task-oriented conversations present several challenges that limit their effectiveness. Firstly, they consist of distinct components – intents, slots, and utterances – that impact the similarity and overlap between conversations. For instance, users can have different objectives (e.g., booking travel vs. product returns), or even have the same intents but provide different levels of slot information (Ruane et al., 2018). Additionally, information is typically provided over multiple conversation turns, and each turn could involve multiple user intents and slots. Finally, the same set of tasks can be expressed using numerous possible utterances by users, depending on their

choice of phrasing, order of sentences, use of colloquialisms, introducing digressions, etc. (Guichard et al., 2019). Hence, relying solely on distance based similarity of word-embeddings would adversely impact performance.

In this work, we present *TaskSim*, a novel similarity metric designed for task-oriented conversations to address the above challenges. *TaskSim* represents the structure of conversations as distributions over the different task-oriented components and combines the geometry of the distributions with optimal transport to measure the similarity between conversations. Our approach is inspired by prior work in topic modelling (Kusner et al., 2015; Yurochkin et al., 2019) that have shown the effectiveness of comparing the structure of distributions, albeit for different settings. We evaluate *TaskSim* on a benchmark task-oriented conversation dataset and demonstrate its effectiveness while present detailed use-cases illustrating its improvement over existing approaches.

2 Illustrative Example

Conversations often constitute multiple instructions executed in different orders by different users. Figure 2 shows users 1 and 2 having similar conversation about making bookings for a trip but with re-ordered tasks. An ideal metric to measure conversational similarity should be able to identify that the overall goal of these conversations is the same.

3 Similarity Metric for Task-Oriented Conversations

3.1 Definition

A task-oriented conversational system supports a pre-defined set of user intents $\mathcal{I}$ and their corresponding slots or parameters $\mathcal{S}$. Each conversation C_i consists of a multi-turn sequence of utterances U_i between the user and the system or agent, a subset of active intents, and slot-value information provided by the user (i.e.) $I_i \subseteq \mathcal{I}$ and $S_i \subseteq \mathcal{S}$ (Figure 1). Our objective is to compute the similarity between task-oriented conversations, given their components (i.e.) utterances, intents, and slot information.

3.2 Approach

The key to *TaskSim* is to measure similarity between task-oriented conversations as a function of the distance between their component-wise distributions. For each component, our goal is to represent its distribution over every conversation and to then determine the cost of transforming or transporting the cumulative component-wise distributions of one conversation to another.

Intuitively, conversations with similar intents, slot information, and analogous language would reflect similar distributions, and hence a lower cost of transportation (i.e.) higher similarity. However, any differences in their components would incur a larger cost, and hence reflect a lower similarity.

We begin by generating embeddings for all the intents and slots within the ontology. We do so on the masked conversations: the slot values in every conversation are masked with their corresponding slot name from the ontology. This is to ensure that entities representing slot values do not incorrectly bias or ambiguate the similarity of the embeddings (Gladkova and Drozd, 2016; Shi et al., 2018). For instance, the embedding similarity between the unrelated utterances - *"I want a ticket to the Big Apple"* and *"I want a ticket to the Apple conference"*, could be unduly influenced by the word 'Apple', but masking with their appropriate slot names (e.g., <arrival_city> and <product_name>), resolves this possibility. We denote Δ_U^l as the simplex of utterance embeddings of a conversation.

We then compute probability simplexes $\Delta_{\mathcal{I}}^n$, $\Delta_{\mathcal{S}}^m$ for each conversation over the set of intents $\mathcal{I}$ and slots $\mathcal{S}$ –

$$\Delta_K^n = \{p_i \in \mathbb{R}^{n+1} \mid \sum_{i=0}^{n} p_i = 1 \ , \ p_i \geq 0 \ \forall i \in |K|\}$$

where each p_i reflects the frequency of occurrence of intents and slots over the utterances. For example, $\Delta_{\mathcal{I}}^n$ for conversation C_i represents the probability of all n intents within C_i. We then compute a cost matrix $\mathcal{M}_{i,j}$ for each component, that represents the cost to move between two points (i, j) in its distribution. We compute each entry using the Euclidean distance between the embeddings generated for each component.

Given simplexes $\alpha \in \Delta_K^n$, $\beta \in \Delta_K^m$ and the cost matrix $\mathcal{M}$, the 1-Wasserstein distance (Vallender, 1974) between them is –

$$W_1(p, q) = \min_{\Gamma \in \mathbb{R}^{n \times m}} \sum\nolimits_{i,j} \mathcal{M}_{i,j} \Gamma_{i,j}$$

$$\text{subject to } \sum\nolimits_j \Gamma_{i,j} = \alpha_i \text{ and } \sum\nolimits_i \Gamma_{i,j} = \beta_j$$

where $\mathcal{M}_{i,j} = d(i, j)$ denotes the cost matrix and $d(.,.)$ denotes the distance between the distributions. We then define the similarity (*TaskSim*)

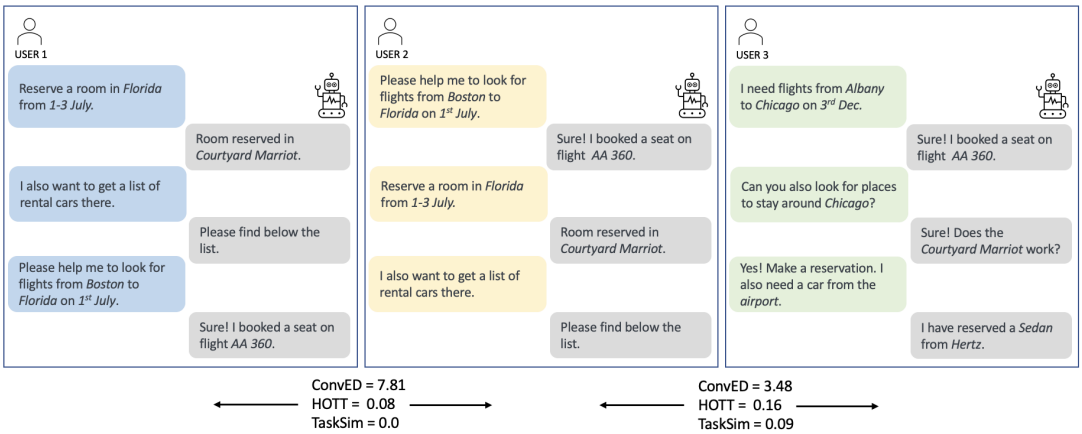

Figure 2: Sample conversations to showcase *TaskSim* similarity values versus other benchmarks

between two task-oriented conversations C_1 and C_2 as the weighted sum of the W_1 distances between their respective components –

$$\text{TaskSim}(C_1, C_2) = \sum \gamma W_1(C_1^{\oplus}, C_2^{\oplus}) \quad (1)$$

where $C_i^{\oplus} = \{U_i, I_i, S_i\}$ represents the conversation's components (i.e.) utterances, intents, and slots, and γ is a hyperparameter reflecting the influence of each component on the similarity.

4 Experimental Evaluation

4.1 Experimental Setup

We conduct our experiments on an Intel Core i9 with 64GB of RAM. We implement *TaskSim* in Python, leveraging the POT library (Flamary et al., 2021) for the 1-Wasserstein optimal transport distance. γ is set to 2, 1, and 1 for the intent, utterances and slots components, respectively.

Dataset We use SGD (Rastogi et al., 2020), a benchmark dataset of multi-turn task-oriented conversations between users and agents spanning 20 domains like travel, dining, weather, etc. Its 20,000 conversations are annotated with active intents and slot information.

Baselines We compare *TaskSim* to three approaches from the literature. Conversational Edit Distance (ConvED) (Lavi et al., 2021) is a dialogue similarity metric that aligns utterances between conversations and computes the edit distance between their embeddings. Hierarchical Optimal Transport (HOTT) (Yurochkin et al., 2019) computes similarity between documents by modeling their topics using latent Dirichlet allocation (LDA) (Blei et al., 2003), and subsequently uses the 1-Wasserstein distance on the topic and text embeddings. Cosine Similarity (Cosine) measures similarity based on the cosine of the angle between text embeddings.

4.1.1 k-NN classification Results

We perform k-NN classification to demonstrate *TaskSim*'s ability to classify conversations into the correct domain. Table 1 shows the accuracy of the classification (on 1 run). Popular similarity metrics like cosine perform relatively well, but rely only on utterances to measure text similarity. Thus, different phrasing and entity names highly influence the metric even though the conversations are very similar. Topic-based metrics like HOTT produce poor accuracy scores as topic modeling methods like LDA use word-frequency distributions which are unable detect relevant topics from short utterances. The masking of slot values and combination of other dialogue features like intents and slots helped *TaskSim* overcome the shortcomings of existing textual and document based similarity metrics.

	Accuracy
Cosine	0.78
HOTT	0.15
ConvED	0.86
TaskSim	**0.95**

Table 1: Accuracy scores for k-NN classification

4.1.2 Conversational clusters Results

To highlight the effectiveness of our approach, we visualize the conversational clusters formed by k-means clustering over 20 iterations. We set the number of clusters to 24, the total number of domains in the dataset. Figure 3 shows the colored visualization of the clusters. The well-formed distinct clusters demonstrate the ability of *TaskSim* to efficiently identify similar conversations and thus could improve the performance downstream conversational analytics tasks.

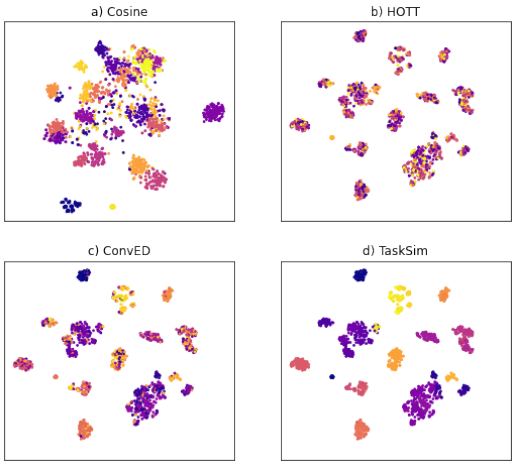

Figure 3: Conversations clustered using k-means and color coded by domain names.

4.1.3 Ablation Study

We also perform an ablation study to investigate the influence of each component in *TaskSim*. The classification accuracy scores (c.f. Table 2) how dialogue specific features like intents and slots aid in improving the distance metric that only uses utterances. We conclude that *TaskSim* components can be used in isolation but provide maximized value when used in combination with slot descriptions and intents.

	Accuracy
TaskSim	0.95
- Utterance	0.88
- Slot	0.93
- Intent	0.94

Table 2: k-NN classification accuracy after removing components of *TaskSim*

4.2 Robustness to Re-ordering

To demonstrate this capability, we perform a perturbation analysis on the SGD dataset wherein 30% of the utterances in each conversation have been reordered and compared with the original one (Table 3). An edit distance based metric, considers a conversation as a sequence of utterances only, which fails to identify similar tasks in a different sequence. *TaskSim* correctly captures the average distance between conversations as 0 since it represents the conversation as a distribution over all the intents, slots and utterances. The use of intents and slot descriptions helps to identify the overall goal and closeness of individual tasks in the conversation.

5 Related Work

Efforts across many natural language tasks including sentiment analysis (Poria et al., 2016), recommendation systems (Magara et al., 2018), and question answering (Sidorov et al., 2015), have relied on using distance-based similarity measures over text embeddings (Wang and Dong, 2020). In addition to these measures, recent work on dialogue similarity have also leveraged conversation structure, where Appel et al. (2018) consider the number of dialogue turns, words, and cycles and use cosine similarity. Similarly, Xu et al. (2019) cluster user-bot dialogues using different distance measures, Enayet and Sukthankar (2022) measure similarity of dialogue sequences using the Hamming distance, and Lavi et al. (2021) align user-bot utterances before using edit-distance to compute similarity.

The use of optimal transport over text distributions has shown promising results in document similarity (Solomon, 2018) resulting in popular metrics like the word mover's distance (WMD) (Kusner et al., 2015) and supervised WMD (Huang et al., 2016). Recently, Yurochkin et al. (2019) used optimal transport over topic models for documents, demonstrating a significant improvement in performance over traditional distance based measures. However, direct application of such approaches to task-oriented dialogues is challenging, due to the unique structure and different components of conversations, as shown in our results.

Approach	**Avg. Distance**
HOTT	0.200
ConvED	4.150
Cosine	0.005
TaskSim	**0.000**

Table 3: Impact of conversational re-ordering on performance of all approaches

6 Limitations and Conclusion

TaskSim was designed specifically for task-oriented conversations; it not only captures semantic similarity between the utterances but also utilizes dialog specific features like intents and slots to identify the overall objective of the conversation. Future work will further validate its effectiveness by carrying out human evaluations to investigate the inclusion of additional dialog features on open domain dialog datasets, more extensive experiments to demonstrate the hyper-parameter optimization methodology to weigh different dialog components in *TaskSim*, and experiments that evaluate the effects of more accurate distance values on downstream tasks like prompt engineering.

References

Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, et al. 2020. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571.

Ana Paula Appel, Paulo Rodrigo Cavalin, Marisa Affonso Vasconcelos, and Claudio Santos Pinhanez. 2018. Combining textual content and structure to improve dialog similarity. *arXiv preprint arXiv:1802.07117*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Ayesha Enayet and Gita Sukthankar. 2022. An analysis of dialogue act sequence similarity across multiple domains. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3122–3130.

Rémi Flamary et al. 2021. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

Jianfeng Gao, Chenyan Xiong, and Paul Bennett. 2020. Recent advances in conversational information retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2421–2424.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 36–42.

Jonathan Guichard, Elayne Ruane, Ross Smith, Dan Bean, and Anthony Ventresque. 2019. Assessing the robustness of conversational agents using paraphrases. In *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)*, pages 55–62. IEEE.

Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. 2016. Supervised word mover's distance. *Advances in neural information processing systems*, 29.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Ofer Lavi, Ella Rabinovich, Segev Shlomov, David Boaz, Inbal Ronen, and Ateret Anaby Tavor. 2021. We've had this conversation before: A novel approach to measuring dialog similarity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1169–1177.

Shuyang Li, Bodhisattwa Prasad Majumder, and Julian McAuley. 2022. Self-supervised bot play for transcript-free conversational recommendation with rationales. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 327–337.

Maake Benard Magara, Sunday O Ojo, and Tranos Zuva. 2018. A comparative analysis of text similarity measures and algorithms in research paper recommender systems. In *2018 conference on information communications technology and society (ICTAS)*, pages 1–5. IEEE.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Federica Bisio. 2016. Sentic lda: Improving on lda with semantic similarity for aspect-based sentiment analysis. In *2016 international joint conference on neural networks (IJCNN)*, pages 4465–4473. IEEE.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Elayne Ruane, Théo Faure, Ross Smith, Dan Bean, Julie Carson-Berndsen, and Anthony Ventresque. 2018. Botest: a framework to test the quality of conversational agents using divergent input examples. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, pages 1–2.

Yong Shi, Yuanchun Zheng, Kun Guo, Wei Li, and Luyao Zhu. 2018. Word similarity fails in multiple sense word embedding. In *International Conference on Computational Science*, pages 489–498. Springer.

Grigori Sidorov, Helena Gómez-Adorno, Ilia Markov, David Pinto, and Nahun Loya. 2015. Computing text similarity using tree edit distance. In *2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC)*, pages 1–4. IEEE.

Justin Solomon. 2018. Optimal transport on discrete domains. *AMS Short Course on Discrete Differential Geometry*.

SS Vallender. 1974. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786.

Jiapeng Wang and Yihong Dong. 2020. Measurement of text similarity: a survey. *Information*, 11(9):421.

Luxun Xu, Vagelis Hristidis, and Nhat XT Le. 2019. Clustering-based summarization of transactional chatbot logs. In *2019 IEEE International Conference on Humanized Computing and Communication (HCC)*, pages 60–67. IEEE.

Avi Yaeli, Segev Shlomov, Alon Oved, Sergey Zeltyn, and Nir Mashkif. 2022. Recommending next best skill in conversational robotic process automation. In *International Conference on Business Process Management*, pages 215–230. Springer.

Xin Ye, Hui Shen, Xiao Ma, Razvan Bunescu, and Chang Liu. 2016. From word embeddings to document similarities for improved information retrieval in software engineering. In *Proceedings of the 38th international conference on software engineering*, pages 404–415.

Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin M Solomon. 2019. Hierarchical optimal transport for document representation. *Advances in Neural Information Processing Systems*, 32.

Guangyou Zhou, Tingting He, Jun Zhao, and Po Hu. 2015. Learning continuous word embedding with metadata for question retrieval in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 250–259.