Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3122–3130
Marseille, 20-25 June 2022

An Analysis of Dialogue Act Sequence Similarity Across Multiple Domains

Ayesha Enayet and Gita Sukthankar
Department of Computer Science
University of Central Florida
ayeshaenayet@knights.ucf.edu, gitars@eecs.ucf.edu

Abstract

This paper presents an analysis of how dialogue act sequences vary across different datasets in order to anticipate the potential degradation in the performance of learned models during domain adaptation. We hypothesize the following: 1) dialogue sequences from related domains will exhibit similar n-gram frequency distributions 2) this similarity can be expressed by measuring the average Hamming distance between subsequences drawn from different datasets. Our experiments confirm that when dialogue acts sequences from two datasets are dissimilar they lie further away in embedding space, making it possible to train a classifier to discriminate between them even when the datasets are corrupted with noise. We present results from eight different datasets: SwDA, AMI (DialSum), GitHub, Hate Speech, Teams, Diplomacy Betrayal, SAMsum, and Military (Army). Our datasets were collected from many types of human communication including strategic planning, informal discussion, and social media exchanges. Our methodology provides intuition on the generalizability of dialogue models trained on different datasets. Based on our analysis, it is problematic to assume that machine learning models trained on one type of discourse will generalize well to other settings, due to contextual differences.

Keywords: dialogue acts, sequence analysis, generalizability, domain shift

1. Introduction

Transfer learning is commonly used in natural language processing to compensate for paucity of data; a machine learning model can often be trained on a single large source dataset and then fine-tuned for smaller target datasets. Unfortunately many machine learning models perform poorly when exposed to *domain shifts*, distributional differences between source and target datasets. Studies have shown that, unlike machine learning algorithms, humans are more robust to these natural distribution shifts (Miller et al., 2020).

Our research tackles the problem of learning models for discourse analysis that generalize across different communication settings. Discourse is often represented as a sequence of dialogue acts (DAs) where each DA represents the functional purpose of the utterance in the conversation (e.g., statement, question, agreement). Dialogue modeling systems not only analyze the content of the utterance, but also the context of neighboring dialogue acts to track conversational state; for instance, agreement dialogue acts often follow questions. Due to differences in the linguistic features of training and test data, natural distribution shifts may occur (Kulkarni et al., 2020). In dialogue models that rely on the context of utterances, we hypothesize that differences in DA patterns will affect model performance.

This paper presents a methodology for predicting the potential degradation in the performance of learned models during domain adaptation. Our analysis shows that dialogue sequences from related domains possess similar n-gram frequency distributions. This similarity can be quantified by measuring the average Hamming distance between subsequences drawn from different datasets. We analyze the similarity of the dialogue acts across eight different datasets: SwDA, AMI (DialSum), GitHub, Hate Speech, Teams, Diplomacy Betrayal, SAMsum, and Military (Army). These datasets represent many types of discourse including collaboration, formal discussion, strategic planning, and social media exchanges. Rather than evaluating performance on a specific dialogue modeling task, we evaluate the suitability of embeddings learned from DA sequences for discriminating between discourse from different datasets. Our experiments demonstrate that when dialogue acts sequences from two datasets are dissimilar they lie further away in embedding space, making it possible to train a classifier that is robust to data perturbations, such as random deletion and tag swapping. Our objective is to provide intuition on the transferability of learned models that utilize dialogue act patterns to make predictions; our research findings have implications for many critical applications including conversational agents, question answering systems, role identification, and speech recognition.

2. Related Work

Dialogue act sequences have been leveraged for a variety of NLP tasks such as coreference resolution (Agrawal et al., 2017), misunderstanding detection (Aberdeen and Ferro, 2003), abstractive summarization (Goo and Chen, 2018), discourse chunking (Midgley and MacNish, 2003), information need classification (Frummet et al., 2019), and conversational models (Kumar et al., 2018). Example applications include situational-based dialogue management systems (Lee et al., 2006), agenda-based simulators for train-

ing dialogue managers (Schatzmann et al., 2007), semi-automated negotiation (Zhao et al., 2018), and dynamic dialogue selection (Ryan et al., 2016).

Dialogue act classifiers tag each utterance with a label according to a taxonomy of conversational functions. Many dialogue act classification techniques make use of the labels of the surrounding utterances such as the Contextual Dialogue Act classifier (CDAC) (Ahmadvand et al., 2019), n-gram models (Webb et al., 2005; Grau et al., 2004), and unsupervised multimodal feature-based techniques (Ezen-Can et al., 2015). Neural architectures (Tran et al., 2017b) are commonly employed for dialogue act classification including the dual-attention hierarchical RNN (Li et al., 2018) and generative models (Tran et al., 2017a).

However, there is little work on the problem of measuring similarity between two dialogues. Lavi et al. (2021) introduce a method ConvEd to calculate the similarity between two conversations to support the retrieval of relevant customer service interactions for chatbots. ConvEd measures the edit distance between the two conversations by counting the insertion, deletion, and substitution operations required to align the two conversations. Unlike our work, ConvEd measures similarity by calculating an embedding over the original utterances, rather than the dialogue act tags.

Researchers have developed techniques for efficient computation of document similarity (Elsayed et al., 2008), node similarity (Reyhani Hamedani and Kim, 2021), entity resolution (Chen et al., 2019), and query expansion (Liu et al., 2017), Many of the proposed approaches exploit word embeddings for the computation of similarity (Elsayed et al., 2008; Reyhani Hamedani and Kim, 2021; Chen and Di Eugenio, 2013; Liu et al., 2017). We use Doc2Vec (Le and Mikolov, 2014), a variant of Word2Vec (Mikolov et al., 2013), since we are interested in document level (dialogue) embeddings rather than word level. The Distributed Memory Model of Paragraph Vectors (PV-DM) model of Doc2Vec generates embeddings by sampling context windows of user-defined sizes from a paragraph and preserving the most meaningful information contained in the sequences present in those context windows. The next section describes our methodology for quantifying the similarity of dialogue act sequences.

3. Methodology

A DA classifier was used to extract sequences of dialogue acts from sets of dialogues. Our analysis was performed on eight datasets that span a rich cross-section of human social interactions. First we present the frequency distribution of the dialogue act n-grams. Then we introduce our proposed similarity measure for predicting generalizability performance: the percentage of zero Hamming distance subsequences of fixed window size drawn from different datasets.

We contrast this method to one of the most commonly used methods of calculating document similarity, a Doc2Vec embedding. This type of embedding is often used as a basis for other dialogue modeling tasks. We measure the cosine similarity of discourse using the embeddings obtained through Doc2Vec. Then we study how effective the embedding is at discriminating between dialogue instances drawn from different datasets, using a discriminative distance method. Binary classifiers are trained to classify the dataset from a DA sequence represented in the Doc2Vec embedding; using the learned models, we identify the most confusing pairs of datasets for a binary classifier. We show that the most confusing datasets are typically collected within the same communication context and are highly similar according to both the dialogue act n-gram and Hamming distance analysis. These confusing pairs are strong candidates to be compatible domain adaptation source and target tasks. We have made the dataset of dialogue act sequences collected from different communications settings available at `https://github.com/ayeshaEnayet/DAC-USE` (under DomainShift).

3.1. Dialogue Act Classification

First we apply our Universal Sentence Encoder (USE) based DA classification model, trained on the SwDA dataset, to tag the utterances of all the datasets. We use the SwDA-DAMSL tagset available at `https://web.stanford.edu/~jurafsky/ws97/manual.august1.html`. USE is itself trained on a variety of datasets, including discussion forums, and it exhibits a good performance on a variety of NLP tasks (Cer et al., 2018). The code and details for the DA classification model are available at `https://github.com/ayeshaEnayet/DAC-USE`. We selected the USE based model due to its ability to generalize effectively across dialogue (discussion) datasets. The test accuracy of our classification model is 72%, and validation accuracy is 70% which is comparable to most of the DA classification approaches. The DA classifier does not consider surrounding utterances to predict the tag of the current utterance; classification is performed solely on the basis of the information present in the embedding of a single utterance.

The DA classifier takes a sequence of utterances as its input and returns the sequence of DAs, where each DA corresponds to one utterance. Table 1 shows the top three most frequent unigrams, bigrams, trigrams, 4grams, and 5grams of the datasets used in this analysis. There is some overlap in the DA n-grams across all datasets; for instance sequences of sd (statement-non-opinion) are common across all datasets.

3.2. Datasets

Datasets were selected to represent a cross-section of communication domains including social media exchanges, collaboration, formal discussion, telephonic conversation, and strategic dialogues. Some of these

Dataset	Unigrams	Bigrams	Trigrams	4grams	5grams
Teams	(sd),(b),(%)	(sd,sd),(sd,b),(b,sd)	(sd,sd,sd), (sd,sd,b), (sd,b,sd)	(sd,sd,sd, sd), (sd,sd,sd,sd), (sd,sd,sd,b)	(sd,sd,sd,sd, sd), (sd,sd,sd,sd,b), (sd, sd,sd,b,sd)
GitHub	(sd),(sv),(ad)	(sd,sd),(sd,sv),(sv,sd)	(sd,sd,sd), (sv,sd,sd), (sd,sd,ad)	(sd, sd, sd, sd), (sd, sd, sd, ad), (sv, sd, sd, sd)	(sd, sd, sd, sd, sd), (sd, sd, sd, sd, ad), (sd, sv, sd, sd, sd)
Army	(sd),(qy),(%)	(sd,sd),(qy,sd),(sd,qy)	(sd,sd,sd), (sd,sd,qy), (qy,sd,sd)	(sd, sd, sd, sd), (sd, sd, sd, qy), (qy, sd, sd, sd)	(sd, sd, sd, sd, sd), (qy, sd, sd, sd, sd), (sd, sd, sd, sd, qy),
SAMsum	(sd),(sv),(fc)	(sd,sd),(sv,sd),(sd,sv)	(sd,sd,sd), (sd,sv,sd), (sv,sd,sd)	(sd, sd, sd, sd), (sd, sd, sv, sd), (sd, sv, sd, sd)	(sd, sd, sd, sd, sd), (sd, sd, sd, sv, sd), (sd, sd, sv, sd, sd)
Hate Speech	(sd),(sv),(fc)	(sd,sd),(sv,sd),(sd,sv)	(sd,sd,sd), (sd,sv,sd), (sv,sd,sd)	(sd, sd, sd, sd), (sd, sd, sv, sd), (sd, sv, sd, sd)	(sd, sd, sd, sd, sd), (sd, sd, sd, sd, qh), (sd, sd, sd, qh, sd)
SwDA	(sd) (sv)(b)	(sd, sd),(sd, b),(b, sd)	(sd, sd, sd), (sd, sd, b), (sd, b, sd)	(sd, sd, sd, sd), (sd, sd, sd, b), (sd, sd, b, sd)	(sd, sd, sd, sd, sd), (sd, sd, sd, b, sd), (sd, sd, sd, sd, b)
AMI	(sd), (b),(sv)	(sd,sd),(b,sd),(sv, sd)	(sd,sd,sd), (b, sd,sd), (sd, sv, sd)	(sd, sd, sd, sd), (b, sd, sd, sd), (sd, sv, sd, sd)	(sd, sd, sd, sd, sd), (b, sd, sd, sd, sd), (sd, sd, sd, sd, sv)
Diplomacy	(sd),(sv),(qy)	(sd,sd),(sv,sd),(sd,sv)	(sd,sd,sd), (sd,sd,sv), (sv,sd,sd)	(sd, sd, sd, sd), (sv, sd, sd, sd), (sd, sd, sd, sv)	(sd, sd, sd, sd, sd), (sd, sv, sd, sd, sd), (sd, sd, sd, sd, sv)

Table 1: N-gram frequency distribution: top three most frequent unigrams, bigrams, trigrams, 4grams, 5grams of all the datasets. Sequences of sd (statement-nonopinion) are common across all datasets. The most frequent tags in this table are sd: Statement-non-opinion, b: Acknowledge, %: Uninterpretable, sv: Statement-opinion, ad: Action-directive, qy: Yes-No-Question, fc: Conventional-closing, qh: Rhetorical-Questions.

datasets are quite large, but many are too small to support complex machine learning models. Our analysis was performed on a balanced dataset with 50 randomly sampled dialogues selected from each dataset, except for the Military dataset which only has 22 examples. A noisy version of this dataset was also created by randomly deleting and swapping dialogue act labels (see Section 4.1 for details). All the datasets contain dialogue in the English language. Following is a brief description of the datasets that are used for analysis.

1. **SwDA** is one of the most popular public datasets for DA classification. It consists of 1155 human-to-human telephone speech conversations [1]. The dataset is tagged using 42 tags from the SwDA-DAMSL tagset, which is a subset of Dialogue Act Markup in Several Layers (DAMSL) categories (1997). A more detailed description of SwDA-DAMSL is given by Jurafsky et al. (1997). [2].

2. **SAMsum** is a chat dialogue dataset that consists of Messenger, Whatsapp, and WeChat conversations, written and created by linguists. The dataset contains 16,369 dialogues which include 14,732 train, 819 test, and 818 validation dialogues (Gliwa et al., 2019).

3. **DialSum**, a subset of the AMI meeting corpus, contains 24,193 total dialogues, divided into 7,024 train, 400 test, and 400 validation instances. It is a subset of the AMI meeting corpus with the topic descriptions as abstractive summaries. The AMI meeting corpus contains transcriptions of 100 hours of meeting recordings [3].

4. **Teams** contains 124 team dialogues from 62 different teams, playing two different collaborative board games. The length of the dialogues varies from 291 to 2124 utterances (Litman et al., 2016).

5. **GitHub** is an online platform where software developers collaborate to develop code and discuss software related issues. We collected a dataset from 100 different GitHub issues. The sequence of comments from each issue forms one dialogue of the dataset [4]. The length of the dialogues in our GitHub corpus varies from 2 to 207 utterances. Utterances from the GitHub dialogues blend English language words, special symbols, and code

[1] https://github.com/cgpotts/swda

[2] https://web.stanford.edu/~jurafsky/ws97/manual.august1.html

[3] https://github.com/MiuLab/DialSum

[4] https://github.com/ayeshaEnayet/DAC-USE

written in different programming languages. The average length of the dialogues is 19. The number of speakers varies from 2 to 10.

6. **Diplomacy Betrayal** dataset consists of communication between online users playing the Diplomacy strategic board game. The dataset contains games with different outcomes: half of which ended in betrayal and half ended in friendship [5].

7. The **Hate Speech** dataset consists of utterances extracted from the posts of white supremacist forum. The sentences of the posts are annotated to reflect the presence or absence of Hate Speech [6].

8. The **Military** team communication dataset (Kalia et al., 2017) contains 22 chats from 20 chat rooms. The chats are communication from simulation activity (SIMEX). The average number of speakers in this corpus is 15, and the length of the dialogue varies from 55 to 1027 utterances.

3.3. Sequence Similarity

The Hamming distance between two sequences is the number of positions where the sequences have different values. We extract all the possible subsequences of lengths four and five from the output of the DA classifier and calculate the Hamming distance between the sequences from all the datasets. To score each sequence, we increment the count by one for every pair of subsequences possessing a Hamming distance of zero. The similarity score between two dialogues is represented as a percentage. The final similarity score between datasets is quantified by taking the average of the scores.

3.4. Embeddings

Most machine learning models start by learning a lower dimensional representation of the data that can be used by the NLP pipeline. Each discourse is initially represented as a sequence of dialogue acts. Sequences of DAs are treated as documents, with the DAs forming the vocabulary of the document. We apply Doc2Vec (Le and Mikolov, 2014), a technique to learn paragraph vectors, to learn embeddings from these sequences of dialogue acts. The Distributed Memory (DM) model of the Doc2Vec was used because of its ability to generate embeddings by considering the context window of varying sizes, as opposed to Distributed Bag of Word (DBOW) model, which does not consider the context when learning embeddings. Our analysis was performed with the Doc2Vec function from the Gensim library. We use PV-DM with epoch size of 5, negative sampling 5, and window size 5. We then apply both the discriminative distance method and cosine similarity measures to the embeddings.

[5] https://sites.google.com/view/qanta/projects/diplomacy

[6] https://github.com/Vicomtech/hate-speech-dataset

Discriminative Distance: Discriminative distance was used to identify the most confusing dataset pairs. We train a support vector machine (SVM) binary classifier on the embeddings learned from Doc2Vec; its aim is simply to identify the dataset. The most confusing pairs are the ones that have similar embeddings. If the classifier exhibits a high accuracy, it means that the embedded representation is sufficiently distinct to allow the classifier differentiate between the two datasets. We evaluated the SVM with both a linear and non linear (radial basis function) kernel.

Cosine Similarity: Cosine similarity is a measure of similarity between two vectors calculated by taking the cosine of the angles between two embeddings. We measure the cosine similarity between the embeddings of all the datasets that we obtain through Doc2Vec.

4. Experimental Analysis

The datasets can be grouped by communication setting, with some datasets falling into multiple categories. The Teams, GitHub, and Army datasets are collaborative dialogues gathered from team communications. The SAMsum and Hate Speech datasets are social media exchanges. The Diplomacy and Teams datasets were collected from game communication. GitHub also falls under the social media category, but the dialogues in this dataset are more formal and goal-oriented than SAMsum and Hate Speech. SwDA is a telephonic communication dataset composed of non-goal-oriented discussion between two people. Diplomacy and Army are both good examples of strategic planning. The AMI meeting and GitHub datasets are goal-oriented formal discussion. Table 2 provides an overview of our categorization.

Category	Datasets
Teams, GitHub, Army	Collaboration
SAMsum, Hate Speech, GitHub	Social Media
SwDA	Discussion (informal/non-goal-oriented)
Diplomacy, Army	Strategic planning
Diplomacy, Teams	Gameplay
AMI, GitHub	Discussion (formal/goal-oriented)

Table 2: Categorization of datasets.

Table 1 shows the result of our n-gram frequency distribution analysis and gives the top three most frequent unigrams, bigrams, trigrams, 4grams, and 5grams of all the datasets. The most frequent unigram, bigram, trigram, and 4gram in social media dialogues like SAMsum and Hate Speech are the same. Also, the Yes-No question (qy) is one of the major categories in strategic dialogues. The SwDA and AMI both have statement (sd), opinion (sv), and acknowledgment (b)

as frequently occurring categories in the discourse. Uninterpretable (%) is a prominent unigram in social media datasets. GitHub and Diplomacy datasets have bigram sequences in common; this may occur in both datasets because members propose solutions to each other. Statement-non-opinion (sd) and Statement-opinion (sv) are the most frequently occurring tags of formal dialogues (AMI and GitHub). In addition to sv and sd, the most prominent unigram in GitHub is Action-directive (ad) because, in these dialogues, the members suggest a course of actions to the other members to solve problems. Similarly, in AMI corpus Acknowledge (b) is one of the most prominent tags.

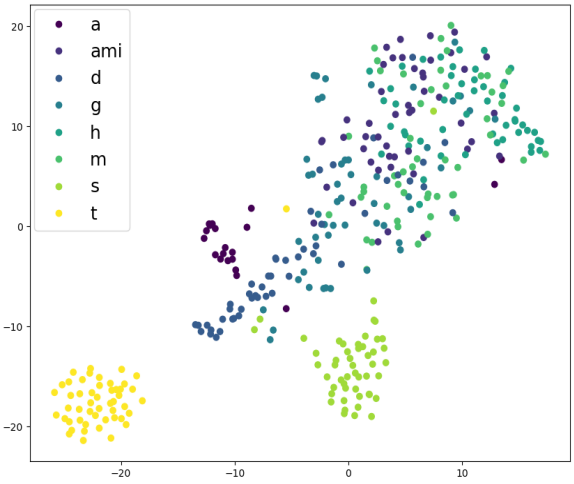


Figure 1: Projection of embeddings of datasets in 2D space. ami: AMI, g: GitHub, d: Diplomacy, t:Teams, h: Hate Speech, s: SwDA, m: SAMsum, a: Military (Army).

Figure 1 shows the distribution of embeddings of all the datasets on a 2D plane. The distribution indicates that SwDA and Teams are clustered separately from other datasets and have unique embeddings. On the other hand, the SAMsum, Hate Speech, and GitHub dataset embeddings (all from the Social Media category) are intermixed and cover a large area. Social media dialogues tend to have a similar dialogue flow. Diplomacy slightly overlaps with GitHub and is near the Military dataset.

Figure 2 shows the classification accuracy of the SVM (with linear kernel) at distinguishing between dialogue act sequences drawn from different datasets. Instances are represented using the embedding illustrated in Figure 1. This shows that SwDA, Military (Army), and Teams are linearly separable from almost all the datasets and exhibit a high classification accuracy. AMI, Hate Speech, GitHub, and SAMsum have high error rates. On the other hand, Diplomacy lies in between highly separable and inseparable datasets. AMI and GitHub, i.e., the formal discussion datasets, showed a significant overlap with four out of seven datasets. The results also indicate that even dialogues

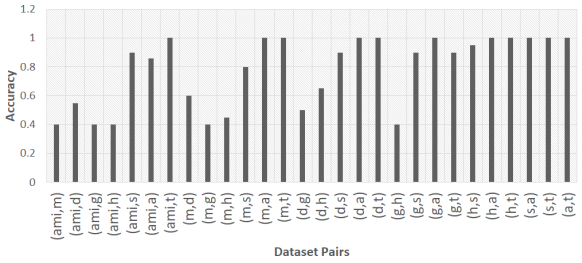


Figure 2: Pairwise classification accuracy using SVM with linear kernel and the Doc2Vec embedding. The classification task is simply to identify the dataset. ami: AMI, g: GitHub, d: Diplomacy, t:Teams, h: Hate Speech, s: SwDA, m: SAMsum, a: Military (Army).

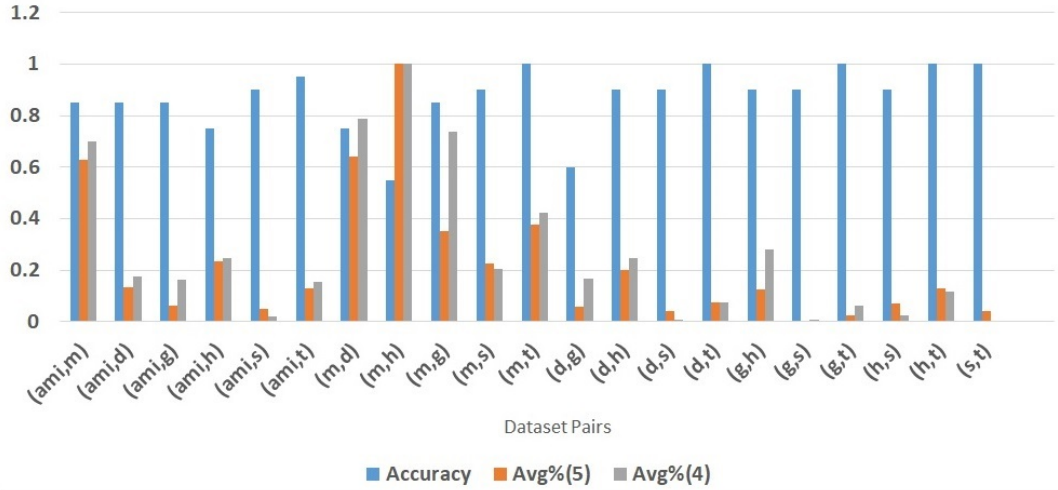
within the same domain may exhibit different communication patterns. The Military and Teams dataset belong to multiple categories but have distinct communication patterns from other datasets.

We validate our ML-based models against the non-ML-based similarity measures. Figure 3 shows the comparison of average percentage similarity between pairs of datasets, calculated using Hamming distance, and binary classification accuracy, using the RBF kernel function. The results show that a high similarity between two datasets leads to low binary classification accuracy. SwDA is one of the standard datasets used for the DA classification task. Yet our results show that SwDA is very different from other datasets, as can be observed in Figure 1, and gives the highest binary classification accuracy when classified against other datasets. SAMsum showed the lowest binary classification accuracy of 55% and 65% when tested with Hate Speech and Diplomacy. SAMsum is one of the datasets which covered a large area in the 2D plane shown in Figure 1; it lies near Hate Speech, GitHub, and Diplomacy.

Table 3 provides an analysis of the cosine similarities of the embeddings. It shows the top two most similar and the least similar datasets for each dataset. The results are consistent with Figure 1 and the binary classification task (Figure 3), showing that SwDA and Teams are two of the least similar datasets. SAMsum and Hate Speech demonstrate a high similarity with almost all the datasets other than SwDA and Teams. SAMsum and Hate Speech are also the datasets that exhibit the poorest binary classification accuracy (see Figure 3) and similar n-gram frequency distributions (see Table 1) with one another. In general, social media datasets exhibit a high degree of similarity.

4.1. Perturbation Analysis

Noise was introduced into the data by performing two perturbations: 1) random deletion and 2) tag swapping. We randomly swap 10% of the tags of each dialogue and generate nine sequences per dialogue. Similarly, we randomly delete 10% tags to generate nine sequences per sequence. This data augmentation strategy



Unscaled Highest Peaks Data (Avg % similarity by Hamming distance)				
Dataset1	Dataset2	Peak value4	Peak value5	Accuracy
SAMSum(m)	Diplomacy(d)	2.608	0.958	0.65
SAMSum(m)	GitHub(g)	2.466	0.548	0.8
AMI (aim)	SAMSum(m)	2.363	0.936	0.85
SAMSum(m)	Hate Speech(h)	3.207	1.459	0.55

Unscaled Lowest Peaks Data (Avg % similarity by Hamming distance)				
Dataset1	Dataset2	Peak value4	Peak value5	Accuracy
GitHub(g)	SwDA(s)	0.411	0.057	0.85
Diplomacy(d)	SwDA(s)	0.411	0.116	0.95
GitHub(g)	Teams(t)	0.566	0.091	0.95
Teams(t)	SwDA(s)	0.393	0.114	100

Figure 3: The trend of binary classification accuracy (for the SVM RBF kernel) vs. average percentage similarity (normalized in the illustration) using the Hamming distance of length 4 and 5 subsequences. Hamming distance similarity predicts poor classification accuracy at the dataset discrimination task. This does not include the results for the Military dataset; its small test set gave 100% accuracy on all the datasets.

Dataset	**Most Similar**	**2nd Most Similar**	**Least Similar**
Army(a)	h(0.4528)	m(0.4494)	s(-0.0526)
AMI(ami)	h(0.4043)	m(0.2880)	s(0.0966)
Diplomacy(d)	h(0.4511)	m(0.4194)	t(-0.0126)
GitHub(g)	h(0.2753)	d(0.2534)	a(0.0285)
Hate(h)	m(0.5281)	a(0.4578)	t(0.1062)
SAMSum(m)	h(0.5352)	a(0.4606)	s(0.0409)
SwDA(s)	h(0.1092)	ami(0.1034)	t(-0.0912)
Teams(t)	ami(0.1464)	h(0.1033)	s(-0.0924)

Table 3: The top two most similar and least similar datasets according to cosine similarity. The cosine similarity for some cases is negative because it is calculated between the embeddings generated through Doc2Vec, not using TF-IDF.

is used to create larger but noisier datasets of dialogue act sequences. We resample the datasets according to the size of the Military dataset and select 140 sequences from each for analysis.

Figure 4 shows the comparison of binary classification accuracy with or without perturbation. The results on the actual dataset vs. the perturbed one show large decreases in the classification accuracy of some of the datasets due to the noise. Altering dialogue act patterns causes the dataset to become similar to some of the other datasets. Figure 5 shows the distribution of synthetic dataset embeddings on a 2D plane. Compared to the embeddings of the original dataset, synthetic dataset embeddings of Diplomacy, Teams, and Army show a slight change in distribution and decreased accuracy with some of the datasets. The formal discussion (AMI and GitHub) perturbed datasets showed a greater decrease in classification accuracy than others.

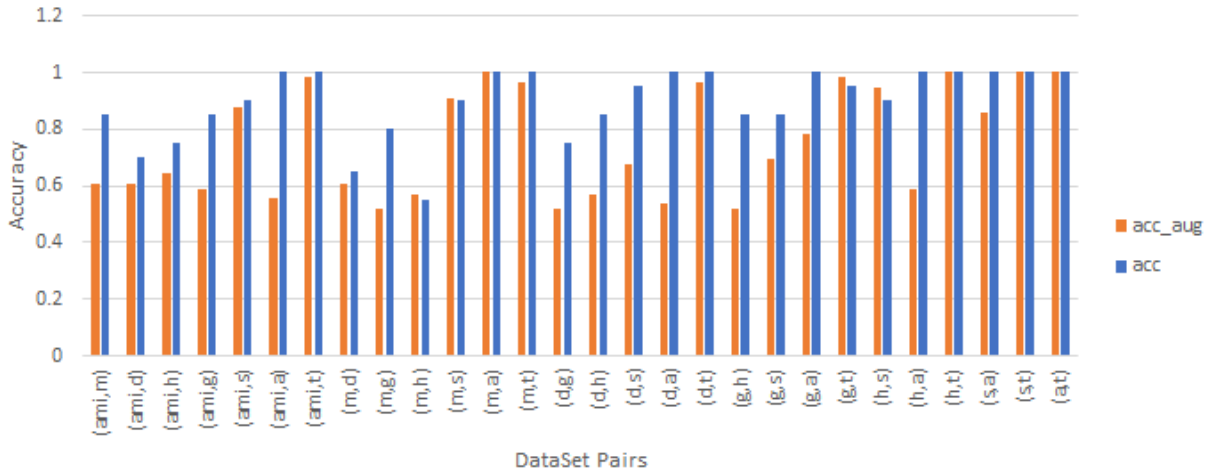


Figure 4: Comparison between the binary classification accuracy of synthetically perturbed data (acc_aug) and actual data (acc). ami: AMI, g: GitHub, d: Diplomacy, t:Teams, h: Hate Speech, s: SwDA, m: SAMsum, a: Military (Army).

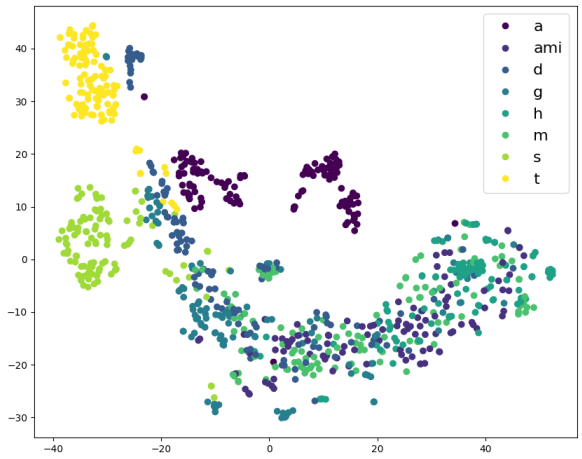


Figure 5: Projection of perturbed dataset embeddings in 2D space. ami: AMI, g: GitHub, d: Diplomacy, t:Teams, h: Hate Speech, s: SwDA, m: SAMsum, a: Military (Army)

Even in the presence of noisy data, the overall distribution of synthetic datasets embeddings, given by Figure 5, is still similar to the embeddings of original datasets (see Figure 1). The learned embedding is clearly robust to slight perturbations.

5. Discussion and Conclusion

This paper presents a dialogue act similarity analysis across multiple communication domains by calculating n-gram frequency distribution, Hamming distance, and the Doc2Vec embedding between dialogue act sequences. It is clear that dialogue act sequences can differ greatly when collected from different communication settings, but even dialogues collected from the same domain can exhibit different communication patterns. The discourse is clearly dependent on the nature and purpose of the conversation. Simple data augmentation techniques like random swap and random deletion tend to alter the dialogue flow such that it becomes more similar to other dialogue categories.

Among all the domains used for the analysis, social media datasets exhibited the highest degree of similarity with one another. Models learned on non-goal oriented discussion do not show potential to generalize well to goal-oriented task specific discussions, and vice versa. One of the most widely used datasets, SwDA, does not exhibit discourse patterns similar to the other datasets used in our analysis. Formal discussions seemed to follow a communication pattern that overlaps with other datasets, and the models learned on these datasets showed a potential to generalize better.

The analysis indicates that the selection of appropriate source and target datasets is equally crucial as developing efficient techniques to achieve generalizability in dialogue and discourse. Based on our analysis, it is problematic to assume that machine learning models trained on one type of discourse will generalize well to other settings, due to contextual differences. We believe our Hamming distance similarity measure can be used to anticipate potential degradation in the performance of learned models during domain adaptation and to select compatible source and target datasets.

6. Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. W911NF-20-1-0008 and ARL STRONG W911NF-21-2-0103. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, ARL, or the University of Central Florida.

7. References

Aberdeen, J. and Ferro, L. (2003). Dialogue patterns and misunderstandings. In *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*.

Agrawal, S., Joshi, A., Ross, J. C., Bhattacharyya, P., and Wabgaonkar, H. M. (2017). Are word embedding and dialogue act class-based features useful for coreference resolution in dialogue? In *Proceedings of PACLING*.

Ahmadvand, A., Choi, J. I., and Agichtein, E. (2019). Contextual dialogue act classification for open-domain conversational agents. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1273–1276.

Allen, J. and Core, M. (1997). Draft of DAMSL: Dialog act markup in several layers.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Chen, L. and Di Eugenio, B. (2013). Multimodality and dialogue act classification in the robohelper project. In *Proceedings of the SIGDIAL Conference*, pages 183–192.

Chen, X., Campero Durand, G., Zoun, R., Broneske, D., Li, Y., and Saake, G. (2019). The best of both worlds: combining hand-tuned and word-embedding-based similarity measures for entity resolution. *BTW*.

Elsayed, T., Lin, J., and Oard, D. W. (2008). Pairwise document similarity in large collections with mapreduce. In *Proceedings of ACL: HLT, Short Papers*, pages 265–268.

Ezen-Can, A., Grafsgaard, J. F., Lester, J. C., and Boyer, K. E. (2015). Classifying student dialogue acts with multimodal learning analytics. In *Proceedings of the International Conference on Learning Analytics and Knowledge*, pages 280–289.

Frummet, A., Elsweiler, D., and Ludwig, B. (2019). Detecting domain-specific information needs in conversational search dialogues. In *Workshop on Natural Language for Artificial Intelligence at AIIA*.

Gliwa, B., Mochol, I., Biesek, M., and Wawer, A. (2019). Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Goo, C.-W. and Chen, Y.-N. (2018). Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742.

Grau, S., Sanchis, E., Castro, M. J., and Vilar, D. (2004). Dialogue act classification using a Bayesian approach. In *Conference of Speech and Computer*.

Kalia, A. K., Buchler, N., DeCostanza, A., and Singh, M. P. (2017). Computing team process measures from the structure and content of broadcast collaborative communications. *IEEE Transactions on Computational Social Systems*, 4(2):26–39.

Kulkarni, R., Hanna, K., and Stanely, J. (2020). NLP generalization for QA tasks.

Kumar, H., Agarwal, A., and Joshi, S. (2018). Dialogue-act-driven conversation model: An experimental study. In *Proceedings of the International Conference on Computational Linguistics*, pages 1246–1256.

Lavi, O., Rabinovich, E., Shlomov, S., Boaz, D., Ronen, I., and Anaby Tavor, A. (2021). We've had this conversation before: A novel approach to measuring dialog similarity. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1169–1177. Association for Computational Linguistics.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196. PMLR.

Lee, C., Jung, S., Eun, J., Jeong, M., and Lee, G. G. (2006). A situation-based dialogue management using dialogue examples. In *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE.

Li, R., Lin, C., Collinson, M., Li, X., and Chen, G. (2018). A dual-attention hierarchical recurrent neural network for dialogue act classification. *arXiv preprint arXiv:1810.09154*.

Litman, D., Paletz, S., Rahimi, Z., Allegretti, S., and Rice, C. (2016). The Teams corpus and entrainment in multi-party spoken dialogues. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1421–1431.

Liu, Q., Huang, H., Lut, J., Gao, Y., and Zhang, G. (2017). Enhanced word embedding similarity measures using fuzzy rules for query expansion. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6.

Midgley, T. D. and MacNish, C. (2003). Automatic dialogue segmentation using discourse chunking. In *Australasian Joint Conference on Artificial Intelligence*, pages 772–782. Springer.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Miller, J., Krauth, K., Recht, B., and Schmidt, L. (2020). The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning*, pages 6905–6916. PMLR.

Reyhani Hamedani, M. and Kim, S.-W. (2021). On investigating both effectiveness and efficiency of embedding methods in task of similarity computation of nodes in graphs. *Applied Sciences*, 11(1):162.

Ryan, J. O., Mateas, M., and Wardrip-Fruin, N. (2016). A lightweight videogame dialogue manager. In *DiGRA/FDG*.

Schatzmann, J., Thomson, B., Weilhammer, K., Ye, H., and Young, S. (2007). Agenda-based user simulation for bootstrapping a POMDP dialogue system. In *Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152.

Tran, Q. H., Haffari, G., and Zukerman, I. (2017a). A generative attentional neural network model for dialogue act classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 524–529.

Tran, Q. H., Zukerman, I., and Haffari, G. (2017b). Preserving distributional information in dialogue act classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2151–2156.

Webb, N., Hepple, M., and Wilks, Y. (2005). Dialogue act classification based on intra-utterance features. In *Proceedings of the AAAI Workshop on Spoken Language Understanding*, volume 4, page 5. Citeseer.

Zhao, R., Romero, O. J., and Rudnicky, A. (2018). Sogo: a social intelligent negotiation dialogue system. In *Proceedings of the International Conference on Intelligent Virtual Agents*, pages 239–246.