LREC 2022

An Analysis of Dialogue Act Sequence Similarity Across Multiple Domains

Ayesha Enayet & Gita Sukthankar

Department of Computer Science, University of Central Florida

Introduction

- 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 dialogue models during domain adaptation.
- Machine learning models perform poorly when exposed to domain shifts, distributional differences between source and target datasets.
- Dialogue modeling systems not only analyze the content of the utterance, but also the context of neighboring dialogue acts to track conversational state.
- Discourse is often represented as a sequence of dialogue acts (DAs).
- In dialogue models that rely on the context of utterances, we hypothesize that differences in DA patterns will affect model performance.
- Example applications include situational-based dialogue management systems[1], semi-automated negotiation[2], and dynamic dialogue selection[3].
- 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).

Table I: Categorization of Datasets

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)

Refrences

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

[2]Zhao, R., Romero, O. J., & Rudnicky, A. (2018, November). SOGO: a social intelligent negotiation dialogue system. In *Proceedings of the 18th International Conference on intelligent virtual agents* (pp. 239-246).

[3]Ryan, J. O., Mateas, M., & Wardrip-Fruin, N. (2016, August). A Lightweight Videogame Dialogue Manager. In *DiGRA/FDG*.

Method and Analysis

A DA classifier was used to extract sequences of dialogue acts from sets of dialogues. We use the following methods to analyze the similarity:

1. Ngram Frequency distribution analysis to identify similar ngrams.
2. We introduce a similarity measure for predicting generalizability performance: the percentage of zero Hamming distance subsequences of fixed window size (4 & 5) drawn from different datasets.'
3. We pass the sequence of DAs through the Doc2Vec to learn embeddings representing the discourse of the dialogues. We then apply the following methods to analyze the similarity between the embeddings.
   - Discriminative Distance: We train a binary classifier to identify the most confusing pairs. Lower the accuracy more the similarity between datasets. We apply both linear and non-linear Kernels.
   - Cosine Similarity.

Figure I: Distribution of Embeddings on 2D space.

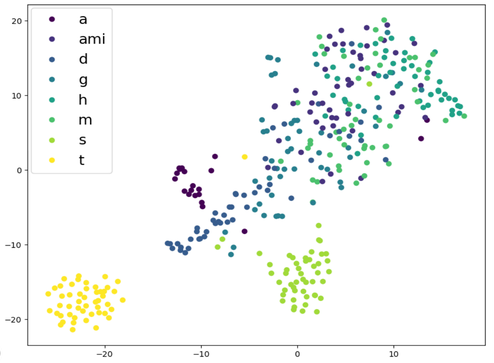
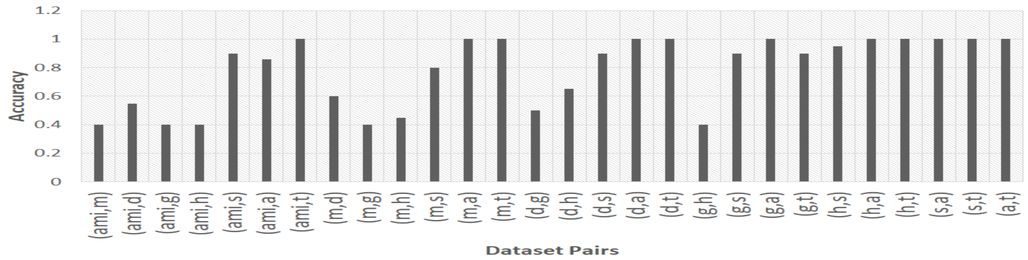
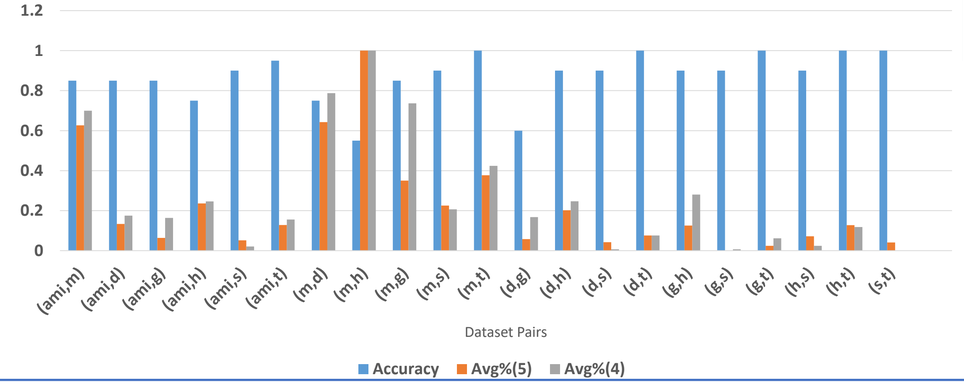


Figure II: Classification Accuracy (Linear Kernel)



Key:
ami:AMI
g:GitHub
d:Diplomacy
t:Teams
s:SwDA
m:SAMSum
h: Hate Speech
a:ARMY

Figure III: Classification Accuracy (Non-Linear Kernel) VS Hamming distance



Ngram Frequency Analysis:

1. Social media dialogues have similar Ngram frequency distribution with Uninterpretable (%) being the most frequent tag.
2. Yes-No question (qy) is one of the major categories in strategic dialogues.
3. Both formal and informal discussion dialogues, SwDA and AMI have statement (sd), opinion (sv), and acknowledgment (b) as frequently occurring categories.
4. In addition to sv and sd, the most prominent unigram in GitHub is Action-directive (ad).

Table II: Cosine Similarity

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)

Conclusion

- Dialogue act sequences can differ greatly when collected from different communication settings.
- The discourse is clearly dependent on the nature and purpose of the conversation.
- 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.
- It could be problematic to assume that machine learning models trained on one type of discourse will generalize well to other settings.

RESEARCH POSTER PRESENTATION DESIGN © 2019
www.PosterPresentations.com