Discovering Domain Specific Dialog Acts:
A Quantitative and Unsupervised Method

Sabina Tomkins
University of California Santa Cruz
satomkin@ucsc.edu

Anbang Xu, Zhe Liu, & Yufan Guo
IBM Research Almaden
{anbangxu,liuzh,guoy}@us.ibm.com

Abstract

We introduce the first unsupervised, end-to-end method that discovers and quantifies the number of dialogue acts in an open domain, with the ability to generate utterance templates for predicted dialog acts. Particularly, it models the sequential behavior of dialog acts as well as the role of emotions in conversation. We evaluate the method on customer service conversations on Twitter from multiple industries. Our method is successful in detecting novel domain specific dialog acts, and in generating appropriate replies to various customer service requests.

1 Introduction

Dialog acts (Stolcke et al., 1998) describe the initial level of speech acts in discourse. They can facilitate conversation modeling and are useful in many conversation applications, such as automatic dialog summarization (Murray et al., 2006; Bhatia et al., 2014) and dialogue systems (Ritter et al., 2010). In accordance with the drastic demand on the service-oriented chatbots, considerable research attention has recently been devoted to automatic dialog act detection within the customer service domain to better characterize and understand the dialog behaviors between the customers and agents (Oraby et al., 2017). Extensive work has been conducted on classifying dialog acts in an effort to establish perfect communication. Many of them are based on supervised learning (Jurafsky et al., 1997; Ji and Bilmes, 2005; Rosset et al., 2008; Oraby et al., 2017; Tavafi et al., 2013; Joty and Hoque, 2016) where numerous training data are required. To save the effort of expert labeling, in this study, we propose an unsupervised method which can automatically detect the number of dialog acts and describe them accordingly. We experiment with data from Twitter, a popular real world outlet for customer service conversations accounting for 80% of similar data in social media (twi, 2016).

2 Method

To model the customer-agent conversations in a finer granularity, we introduce a novel probabilistic graphical model on a sentence level as shown in Figure 1. The model enforces that each sentence depends on the sentence that precedes it, that a response depends on a request, and that requests can influence responses for multiple turns. Dialog acts can be modeled either at the document or sentence level depending on the role of the participant. Customers' emotions are also incorporated into the model as it is highly possible that they affect the responses in a conversation.

Figure 1 illustrates the model for a conversation of two turns, where each turn involves two participants, although it can be generalized to conversations of multiple turns. Each conversation is started by a customer, denoted by pink square nodes (request$_i$) in the figure. An agent replies to the customer, with a tweet containing 1-3 sentences $j \in \{1, 2, 3\}$), with a dialogue act $s_{i,j}$ assigned to each sentence denoted by blue circular nodes. Each reply sentence, represented by a dense vector $d_{i,j}$, is dependent on the corresponding dialogue act as well as the emotional tone of the customer.

Emotion analysis We focused on five primary emotional categories, namely, anger, sadness, joy, disgust, and fear, which are widely accepted in the literature on consumption emotions (Edvardsson, 2005; Devillers et al., 2003). We used [Anonymized] Tone Analyzer API to assign emotion labels to each of the customer tweet, and se-

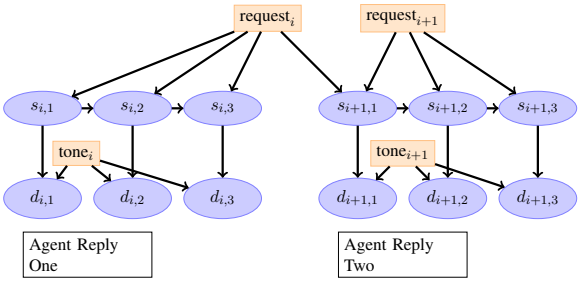

Figure 1: A model of a customer-agent conversation of two turns.

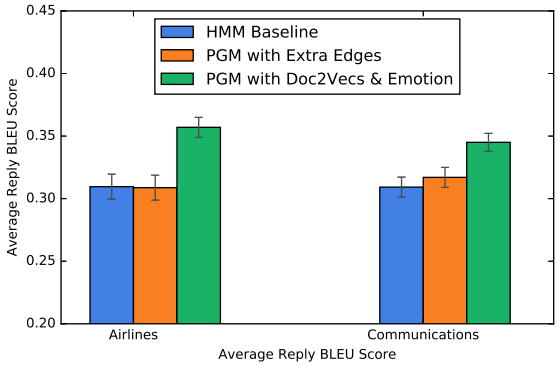

Figure 2: Comparison of our model and baselines.

lected only the dominant emotion.

Dialogue act initialization For faster convergence the dialogue acts were initialized with cluster IDs from Latent Dirichlet Allocation (LDA) (David M. Blei, 2003). Alternative clustering algorithms such as K-Means or Seeded LDA were also investigated and the quality of the generated clusters was approximately the same. Clustering was performed on a tweet level for customers and on a sentence level for agents, as customers' tweets tend to be short, with 1.77 sentences on average, and less complicated than longer agents' tweets with 2.67 sentences on average. TF-IDF was calculated to remove brand/service specific words (e.g. ATT, SIM) to avoid topical bias.

Auto-generated replies Each sentence in a conversation can be represented by a dense, compact vector. This helps to capture the semantic similarity between words, avoid data sparsity, and reduce the number of parameters to be learned. We used the Gensim (Řehůřek and Sojka, 2010) implementation of doc2vec (Le and Mikolov, 2014) to learn sentence embeddings and chose 200 as the vector length. To train the sentence embeddings we use all customer service Tweets in our corpora, and we seed the doc2vec model with Glove(Pennington et al., 2014) word2vecs trained on Twitter[1].

Given an unseen request, an agent's dialogue acts are first predicted. A sentence vector is then drawn from a multivariate Gaussian distribution, based on its dialog act and the observed emotional tone of the customer.

$$d_{i,j} \sim MV(\mu(s,t), \Sigma(s,t) \mid s_{i,j} = s, \text{tone}_i = t).$$

The example reply from the training set whose vector representation has the greatest cosine similarity to the generated vector is returned as a template response. By introducing an empty dialogue act, the model is flexible with replies of various lengths ranging from 1 to 3 sentences.

3 Experiments

The evaluation focuses on the proposed method?s ability to both discover meaningful dialog acts, and to use these acts to generate appropriate replies to unseen requests. We evaluate with two major industries: airlines (@JetBlue, @southwestair, and @DeltaAssist) and telecommunications (@attcares and @TMobileHelp). Over one million tweets were collected using Twitter's Public API between June 2016 and August 2016. User mentions, URL, stopwords, and brand/service specific words were removed from all tweets. Agent tweets were broken into sentences, while customer tweets remained intact. 90% of the data was used in training, while 10% was held out for testing.

The BLEU score (Papineni et al., 2002) was adopted as an evaluation metric for the generated replies and an indirect indicator of the quality of detected dialogue acts. To optimize the number of dialog acts on different datasets, we trained the model 10 times on each dataset, each time with a different number of dialogue acts, ranging from 5 to 20. The optimal number standed out on its own being the one that produced the highest BLEU score of the generated responses. The model is implemented in the python package PYMC3 (Salvatier et al., 2016).

We compare the proposed model with a baseline in terms of the BLEU score. The baseline is an hidden Markov model (HMM), where the dialogue act of the first reply sentence is dependent on the customer request and any subsequent sentence's dialogue act is dependent on the preceding sentence only. As shown in Figure 2, our proposed method outperforms the baseline HMM model with a significantly higher BLEU score.

[1]To initialize the doc2vec model with word2vecs we use: https://github.com/jhlau/doc2vec

Dialog Act	Representative Candidate Sentences
Offering help	We'll see what went wrong.
Comforting	We're working hard to get you on your way to California this morning.
Soliciting information	Have you submitted the Lost Item Report?
Providing information	You can click here for information regarding identification requirements.
Requesting identifiers	What's the flight number?
Empathizing	Sorry to hear you're uncomfortable.

Table 1: Dialog acts for the airline industry.

The optimal performance of our method was achieved when the number of dialog acts equals 6 for the airlines industry (BLEU score of 0.38). The optimal number of dialog acts for telecommunications equals 11 (BLEU score of 0.36). Table 1 lists the representative sentences for each dialogue act detected within the airlines domain.

4 Discussion

We present an end-to-end method for detecting dialog acts. The unsupervised techniques behind our model allows us to dynamically detect dialog acts in any customer service domain, without predefined labels. Our graphical model is targeted at the sequential behavior of dialog acts and the dependence of agent replies on customer requests and emotional tone. Given an unseen request, the embeddings of agent replies can be inferred and their nearest replies in the training data can be returned as a template response. The optimal number of dialogue acts can be automatically determined by optimizing the BLEU score of the generated responses. Using the airline and telecommunication industries as examples, we found that our model produced satisfactory results outperforming the two baseline models.

One advantage of our proposed model is that it enables dynamic dialog act detection within an open domain. A different number of dialog acts have been detected best represented each industry using our model. The airlines industry was best represented by 6 dialog acts, while 11 best captured the nuances of conversations in the telecommunications industry. Dialog acts such as "empathizing", "apologizing", and "soliciting information", were found to be common across brands.

Another advantage of our model is that it can be easily tuned according to the users needs. Take the technical troubleshooting dialog act for example, where across brands agents attempt to guide customers through technical difficulties. This broad dialog act can be furthered refined, to dealing with internet outages, or restarting devices, or interpreting error codes, depending on the product, brand, and user. By providing more targeted data the proposed model has the potential to capture different levels of dialog acts within and across domains.

There are several future directions to pursue. With an end-to-end framework in place, we can now quantitatively evaluate different probabilistic graphical models, clustering algorithms, and document embedding approaches. Document similarity metrics are sensitive to various factors, and a variety of evaluation metrics can be investigated and combined for determining the optimal number of dialogue acts. In the current framework, the generation of agent replies relies on the retrieval of example responses from training data according to the inferred embeddings. A more flexible approach would be to replace sentence embeddings with sequences of word embeddings, and to generate replies through an auto decoder. Additionally, we plan to investigate the proposed model and its extensions with a wider range of data, not only on customer service tweets in dynamic and diverse industries, but also on other domain-specific conversation corpora such as health care and education.

References

2016. Making customer service even better on twitter. https://blog.twitter.com/2016/making-customer-service-even-better/-on-twitter.

Sumit Bhatia, Prakhar Biyani, and Prasenjit Mitra. 2014. Summarizing online forum discussions-can dialog acts of individual messages help? In *EMNLP*.

Andrew Y. Ng Michael I. Jordan David M. Blei. 2003. Latent dirichlet allocation. In *JMLR*.

Laurence Devillers, Lori Lamel, and Ioana Vasilescu. 2003. Emotion detection in task-oriented spoken dialogues. In *International Conference on Multimedia and Expo*.

Bo Edvardsson. 2005. Service quality: beyond cognitive assessment. *Managing Service Quality: An International Journal* 15(2):127–131.

Gang Ji and Jeff A Bilmes. 2005. Dialog act tagging using graphical models. In *ICASSP*.

Shafiq Joty and Enamul Hoque. 2016. Speech act modeling of written asynchronous conversations with task-specific embeddings and conditional structured models. In *ACL*.

D. Jurafsky, R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. Van Ess-Dykema. 1997. Automatic detection of discourse structure for speech recognition and understanding. In *IEEE Workshop on Automatic Speech Recognition and Understanding*.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*.

Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore. 2006. Incorporating speaker and discourse features into speech summarization. In *HLT/NAACL*.

Shereen Oraby, Pritam Gundecha, Jalal Mahmud, Mansurul Bhuiyan, and Rama Akkiraju. 2017. "How May I Help You?": Modeling Twitter Customer Service Conversations Using Fine-Grained Dialogue Acts. In *IUI*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *ACL*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *NACL*.

Sophie Rosset, Delphine Tribout, and Lori Lamel. 2008. Multi-level information and automatic dialog act detection in humanhuman spoken dialogs.

John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. 2016. Probabilistic programming in python using pymc3. *PeerJ Computer Science* .

Andreas Stolcke, Elizabeth Shriberg, Rebecca Bates, Noah Coccaro, Daniel Jurafsky, Rachel Martin, Marie Meteer, Klaus Ries, Paul Taylor, Carol Van Ess-Dykema, et al. 1998. Dialog act modeling for conversational speech. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*.

Maryam Tavafi, Yashar Mehdad, Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2013. Dialogue act recognition in synchronous and asynchronous conversations. In *SIGDIAL*.