Developing a Dataset for Classifying Intents and Sentiments from Judicial Conversations

Palash Nandi[†], **Pinaki Karkun**[†], **Chitra Maji**[†], **Adrija Karmakar**[‡§]
Protyush Jana[‡§], **Arunima Roy**[¶§] and **Dipankar Das**[†]
{sondhanil1, pinaki.karkun, chitra.maji2308, adrijakarmakar2, protyush4711,
arunimaroy1111, dipankar.dipnil2005}@gmail.com
[†]Jadavpur University, Kolkata, West Bengal
[‡]Maulana Abul Kalam Azad University of Technology, West Bengal
[¶]University of Engineering and Management, Kolkata, West Bengal

Abstract

In the present work, we developed a dataset annotated with intents and sentiments at the utterance level. The dataset consists of 430 legal conversations between the user and automated assistant with a total of 2854 utterances (user: 1440, assistant: 1414). The intent annotation follows an ontology provided by experts whereas the sentiment of each user utterance has been evaluated on a scale of -5 to +5. The motivation for including sentiment along with intent was to aid in the generation of an appropriate response. We explored different machine learning (ML) and deep learning (DL) models to accomplish two major tasks: Intent Classification (IC) and Sentiment Classification (SC) to evaluate the usability of the dataset. The results and outcomes were satisfactory for both tasks.

Keywords: intent, sentiment, classification, judicial dataset

1 Introduction

Consultation with a legal expert turns into a necessity to overcome legal issues which can be time-consuming as well as economically challenging. Moreover, serving a large pool of clients simultaneously can be a tiresome job for a legal consultant. A conversational assistant that is able to analyze the client's perspective and suggest accordingly, can be a solution to it. To the best of our knowledge, any large corpus in the legal context is not available to train such an assistant. Thus, we present a conversational dataset in the legal context annotated with intents and sentiments at the utterance level. We also conducted a comparative study of different ML and DL models on the task of IC and SC for assessing usability. The dataset consists of 430 legal conversations with a total of 2854 utterances (user: 1440, assistant: 1414). Each user utterance may fall under multiple intent class out of 29 predefined classes proposed by experts and marked with a sentiment score within a range of -5 to +5 based on annotators' perception. We also carried out a comparative study and error analysis for different models for both intent and sentiment classification. In the case of Intent Classification (IC), Rasa DIET achieves the highest precision of 0.896, recall 0.944, F1-score 0.921, respectively, and outperforms other models. Besides, for Sentiment Classification (SC), RNN performs better in all cases of non-sampling, undersampling, and oversampling in comparison to all other models.

2 Related Work

IC and SC have been in the interest of researchers for a long time. In initial days, lexicon (Kang and Kim, 2003) (Lee et al., 2005), statistical (Liu et al., 2006) or rule (Jansen et al., 2008) based models were considered. In the next decade, authors applied neural models for the same purpose. (Xu and Sarikaya, 2013) had used CNN followed by triangular CRF, (Mesnil et al., 2013) used bi-directional RNN followed by basic CRF where as (Yao et al., 2014) used a modified deep LSTM followed by CRF and softmax for better understanding of the context. (Qin et al., 2019) opted for self-attentive encoder to produce context-aware representation which extracts and summarizes features for IC at sentence and the token level. Recently (Chen et al., 2019) have fine tuned a BERT model for both IC and SC task.

3 Dataset

A total of 2854 utterances were collected [2] from an online legal forum [3] . The raw data was in the form of a sequence of user-posted legal issues and corresponding advice from legal experts (in Indian

§ Equal contribution in the research work as a part of an internship at Jadavpur University.

[2] https://www.crummy.com/software/BeautifulSoup

[3] https://www.kaanoon.com/

Speaker	Statement
User	My husband is abusing me for years.
Legal Expert 1	1. File for divorce 2. Apply for maintenance.
Legal Expert 2	1. Make a police complaint 2. Send him legal notices 3. File divorce (optional)

Table 1: Sample of the scrapped corpus

legal context). Table 1 represents a sample of the raw scrapped corpus.

A total number of 430 different legal cases were collected. Later, the raw dataset was converted into a conversational format. Initially, different advice from different legal experts was analyzed to identify the direction and chronology of the events. Each of the events is represented as a pair of an issue followed by corresponding legal advice. Each of the important pairs was concatenated to form different conversational storylines. Finally, the informative ones are chosen to be included in the dataset. Table 2 represents the possible conversations w.r.t the raw text of Table 1 but only the third conversation was considered suitable.

The dataset consists of 29 intents proposed by experts and each user's utterance is tagged with a sentiment score between -5 to +5. In case of utterances of the agent, the annotation has been limited to intent only.

Id	Conversation
1	**User**: My husband is abusing me for years. **Bot**: You can file for divorce.
2	**User**: My husband is abusing me for years. **Bot**: File for divorce or opt for mutual settlement.
3	**User**: My husband is abusing me for years. **Bot**: File a written complaint at the police station. **User**: We have tried to solve this mutually but failed. **Bot**: Then file a divorce case on the ground of mental cruelty. **User**: But how will I survive if i divorce him? **Bot**: File a maintenance case too.

Table 2: The possible conversations flow w.r.t. the raw text mentioned in Table 1

4 System Description

The system is assessed with two tasks - IC and SC. Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), and Logistic Regression (LR) were employed for both IC and SC whereas Stochastic Gradient Descent Classifier (SGDC), Multi-class BERT (MBERT), Rasa DIET (RDIET) [4] were used for only IC. Lexicon based model (LBM), Random Forest (RF), Convolution Neural Network (CNN), and Recurrent Neural Network (RNN) were used for SC task exclusively.

5 Experimental Results & Observations

Table 3 represents the experimental outcomes of IC. As observed, MNB performs lowest with F1-score of 0.17 for IC. SGDC, SVM, and LR perform similarly but MBERT and RDIET outperform rest of the models. The RDIET or MBERT uses transformer-based approach that aids in better performance.

	Precision	Recall	F1-score
MNB	0.15	0.26	0.17
SGDC	0.38	0.38	0.37
SVM	0.36	0.42	0.37
LR	0.38	0.42	0.39
MBRT	0.59	0.49	0.53
RDIET	**0.89**	**0.94**	**0.92**

Table 3: Experimental result for intent classification models

For SC, the presence of neutral sentiment is highest followed by negative and positive. To eliminate the bias, a separate study was done on undersampled and over-sampled data along with the original one. In all of the cases, CNN is able to score similar to RNN but RNN performs best.

6 Conclusions

This paper aimed to develop a conversational dataset in the legal domain and investigate the usability through IC and SC. As observed, transformer-based models perform best because of better contextual understanding. In the future, we will undoubtedly focus on increasing the amount of training data (including devanagari and code-mixed regional Indian languages) and explore other transformer-based models.

[4] https://rasa.com/

References

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

Bernard J Jansen, Danielle L Booth, and Amanda Spink. 2008. Determining the informational, navigational, and transactional intent of web queries. *Information Processing & Management*, 44(3):1251–1266.

In-Ho Kang and GilChang Kim. 2003. Query type classification for web document retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 64–71.

Uichin Lee, Zhenyu Liu, and Junghoo Cho. 2005. Automatic identification of user goals in web search. In *Proceedings of the 14th international conference on World Wide Web*, pages 391–400.

Yiqun Liu, Min Zhang, Liyun Ru, and Shaoping Ma. 2006. Automatic query type identification based on click through information. In *Asia information retrieval symposium*, pages 593–600. Springer.

Grégoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio. 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Interspeech*, pages 3771–3775.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. *arXiv preprint arXiv:1909.02188*.

Puyang Xu and Ruhi Sarikaya. 2013. Convolutional neural network based triangular crf for joint intent detection and slot filling. In *2013 ieee workshop on automatic speech recognition and understanding*, pages 78–83. IEEE.

Kaisheng Yao, Baolin Peng, Yu Zhang, Dong Yu, Geoffrey Zweig, and Yangyang Shi. 2014. Spoken language understanding using long short-term memory neural networks. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 189–194. IEEE.