Dialog Act Classification for Vietnamese Spoken Text

Thi-Lan Ngo, Thi Bich Ngoc Doan, Thi Lan Phuong Ngo

Abstract— **Systems, which use the conversational interface to interact with users such as chat-bots, virtual personal assistants, recommendation systems and automatic customer care systems and so on, are getting popular in our life. A significant challenge in designing and building those systems is how to effectively determine the user intents from the user's speech interactions. In particular, determining dialog act is the first step in determining user intent. Dialog act recognition has widely studied in many different languages but in Vietnamese, there are few studies. In this paper, we present an attempt on dialog act recognition for Vietnamese conversational text. We adopt a machine learning approach by using maximum entropy model on Vietnamese conversational dataset labelled dialog act based on ISO 24617-2 standard. The achieved result is 70.80% that satisfy for practical applications.**

Index Terms— **Dialog act, ISO 24617-2 standard, spoken text understanding, Vietnamese language processing, dialog system, Vietnamese spoken text.**

I. Introduction

The user interacts with the machine through an interface which can be a console, a graphical interface or conversation interface. Thanks to increasing in voice recognition and natural language processing technology, the applications using a conversational interface such as Virtual assistants (Apple Siri, Microsoft Cortana, Google Assistant, Amazon Alexa or Microsoft XiaoIce [1]) and chat-bot are more and more popular. A conversation interface is an interface for simulating a conversation with a real human. Figure 1 describes the general structure of a conversation interface. First, input speech is converted to plain text by an Automatic Speech Recognition (ASR). The text is analyzed by a Natural Language Understanding module (NLU). The semantic information from the NLU module is analyzed by the dialog manager component which uses the history and state of the dialogue to defines the content of the next utterance and the behaviour of the dialogue system. Output generator module is a natural language generation. Finally, the output is rendered using Text to speech Synthesis. In which, User Intent Determining is one of important part in NLU and Dialog Act (DA) recognition is the first step to recognize user intent. DA also known as speech act or communicative acts [2, 3], represent a speaker's intention.

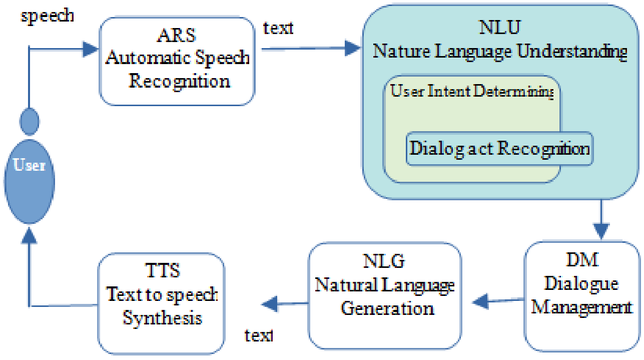

Figure 1. The general structure of a conversational interface. .

Over the years, DAs have been investigated by the research community [4] and have been applied successfully to many tasks [5, 6, 7]. DAs also have been researched in many languages such as Japanese [8, 9], Czech [10], Korean [11, 12, 20], Arabic [13] and so on. However, there are a very limited amount of studies on automatic DA recognition in Vietnamese. With the increasing popularity of systems using a conversational interface for Vietnamese, we are facing the challenge of creating effectively dialog act recognition and annotated data to support the development of dialog modelling systems. In this paper, we present an attempt on annotated data based on ISO standard 24617-2 and dialog act classification for Vietnamese conversation language. We build a classification model using Maximum entropy method to conduct our experiment.

II. Related Work

A. Dialogue Act Annotation Schemes

There are some well-known dialog act annotation schemes such as MapTask[14], Discourse Annotation and Markup System of Labeling (DAMSL) [15] tag-set for the SWDA [16] corpus, Verbmobil [17]. These schemes are very specific to the described scenario, some of its DAs do not scale well to generic and inadequate when applying for non-task-based interaction. Dynamic Interpretation Theory (DIT) [18] is the theoretical foundation for a task-independent DA and domain-independent taxonomy. It was designed to capture all human behaviours during conversations. DIT has been expanded DIT++ to provide a unique and universally recognized standard for DA annotation [19]. Recently, its last version was accepted as an international ISO standard for DA annotation (ISO 24617-2).

ISO 24617-2 defines communicative functions in generic and multiple dimensions intending to remove ambiguities between various aspects of the communication and

Thi-Lan Ngo, Department of Computer Science, University of Information and Communication Technology, Thai Nguyen, Vietnam, +84943870272

Thi Bich Ngoc Doan, Department of Computer Science, University of Information and Communication Technology, Thai Nguyen, Vietnam, +84917326097

Thi Lan Phuong Ngo, Department of Computer Science, University of Information and Communication Technology, Thai Nguyen, Vietnam, +84975272359

overlapping between DAs. So that it is suitable for domain and task independence. It is also befitting for mapping virtually any kind of conversation. Furthermore, using the multidimensional aspect and hierarchical taxonomy make it extensible and potentially adaptable to specific conversational sets. Moreover, the differences in annotation schemes and datasets used in different studies lead the difficult in the mentioning and comparing DA classification approaches. ISO 24617-2 contributed to solve this problem.

B. Dialogue Act Tagging

The automatic dialog acts recognition has been studied by various machine learning techniques such as HMM [4], neural networks [11, 20] or CRF [21], and as a multi-class classification problem using for example SVM [22]. The dialog act classification is a sequence labelling problem. As mentioned, these researches used different DA annotation schemes and datasets. So these are hardly compatible.

In addition, there are a few available training data for the ISO standard. Currently, there are available corpora with ISO 24617-2 annotation (DialogBank [23], ViDa [24]). However, they are too small to be used to train classifiers and imbalanced for the distribution of various DA dimensions. In this work, we labelled dialog act according to ISO 24617-2 for the utterance in a conversation includes 4 people.

III. Experiment

A. Data Building

Our building data process includes six steps which are described in Figure 2.

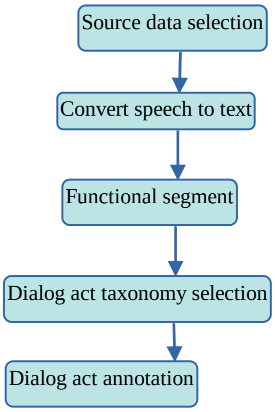

Figure 2. Data building process.

The first step, we selected videos from Gameshow "You want to date" on Vietnamese television to build our data. There are four main characters in the video includes two MCs and two players. Speech in the video is converted into a text by using an auto-sub package () which is an automated voice recognition utility and creates subtitles. It requests the Google Speech API to create records for that conversation automatically to create subtitles. The texts obtained from Auto-sub package have an accuracy of 55%. So, before labelling texts, we revise the texts and edit it manually.

In ISO standard 24617-2, dialog act labels are assigned at "functional segment" level. Functional segment (FS) is "minimum stretching of communication behaviour with one or more communication functions". Therefore, in this study, we split the texts (turn) in the conversations into FSs before labelling dialog act. Examples of turns, FSs and dialog act labels are shown in Table 1. A turn "please woa" is segmented into 2 FS: (fs1) "please" and (fs2) "woa". Another turn "My name is Pham Quang Huy I was born in 1994 I work in Gia Lai my family in Cao Phong town my hometown is a beautiful town and near Hanoi" was divided into 5 Fss: (fs1) "My name is Pham Quang Huy"; (fs2) "I was born in 1994"; (fs3) "I work in Gia Lai"; (fs4) "my family in Cao Phong town"; (fs5) "my hometown is a beautiful town and near Hanoi".

Table 1. Examples about turns, functional segment and dialog act label in the dataset

Turn	FS	Label
Mời bạn woa (please woa)	Mời bạn (please)	direction
	woa	none
Em tên là phạm quang huy em sinh năm 1994 em làm tại Gia Lai gia đình em hiện ở thị xã Cao Phong quê em là thị xã đẹp ở gần hà nội (My name is Pham Quang Huy I was born in 1994 I work in Gia Lai, my family is now in Cao Phong town, my hometown is a beautiful town and near Hanoi)	Em tên là phạm quang huy (My nema is Pham Quang Huy)	inform
	em sinh năm 1994 (I was born in 1994)	inform
	em làm tại Gia Lai (I work in Gia Lai)	inform
	gia đình em hiện ở thị xã Cao Phong (my family is now in Cao Phong town)	inform
	quê em là thị xã đẹp ở gần hà nội (my hometown is a beautiful town and near Hanoi)	inform

The set of labels is selected from the set of labels in ISO 24617-2 and focus on the "Task" dimension, including:

- Question: speaker wants to know the information about something.
- Inform: speaker wants to "provide the addressee certain information which he believes the addressee not to know or not to be aware of, and he believes to be correct".
- Directive: speaker wants the listener to consider a certain action which he might carry out and he is potentially wanting listener do it. For examples requests, orders, instructions, suggestions, etc.
- Commissive: the speaker is committing to perform a certain action in a certain manner or with a certain frequency, possibly dependent on certain conditions, and possibly dependent on listener's consent such as promise, offer, accept a proposal, accept or reject suggestions, accept or reject requests…

- none: a label for FS which does not belong to four labels above.

The dataset contains 3266 FSs labelled dialog act. The distribution of dialog acts in the dataset is presented in Figure 3. Data is imbalance with inform 38.59% , none 37.17%, question 15.34%, direction 8.18% and commissive 0.71%.

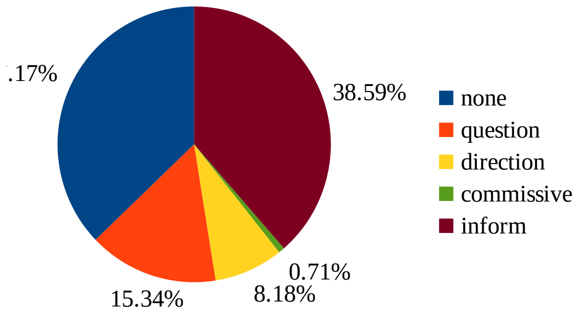

Figure 3. The distribution in the dataset.

B. *Dialog Act Classification Model Using MaxEnt*

To perform our experiment, we approach machine learning technology by using Maxent method. MaxEnt is a short name of maximum entropy method. Its principle is that the most appropriate distribution to model a given dataset is the one with highest entropy among all those that satisfy the constraints of our prior knowledge. Input data is D = {(x_1,y_1), (x_2,y_2), ...(x_n,y_n), in which x_i is a input sentence (in this work input sentence is a FS), y_i is a dialog act label corresponding x_i. There is a distribution which is the maximal entropy solution. This distribution is unique and has form as follow:

$$Z=\sum_{n}\exp(-\sum_{k}\lambda_k f_k(n))$$

where Z is the normalization value, f_i is features of the model, λ_k are appropriate real values usually found by Lagrange multipliers, λ_i are described as a power-law distribution. We need to find λ_i's value. Here, we use the features are n-gram (n=3).

C. *Evaluation Results*

To evaluate the dialog act classification model, we use precision, recall and f_1-score measure. The precision implies the number of correct classifications that are influenced by the incorrect classification. The precision of the c_i class is defined in the following equation:

$$Precision=\frac{a}{b}$$

The recall is the number of correct classifications that are considered with the number of missed entries. The recall of c_i class is count as following:

$$Recall=\frac{a}{c}$$

where:

a is the number of c_i class that the model predicts true and match the label in real.

b is the number of c_i class that the model predicts true.

c is the total number of c_i class that it is true in real (human annotate true in the dataset)

The F1-score measures the harmonic mean of precision and recall, which serves as a derived effectiveness measurement:

$$F1=\frac{2*(precision*recall)}{(precision+recall)}$$

The data is divided into 2 parts: 80% for training and 20% for testing. We performed experiments with 5 folds. Figure 5 presents the results of the model in 5 folds.

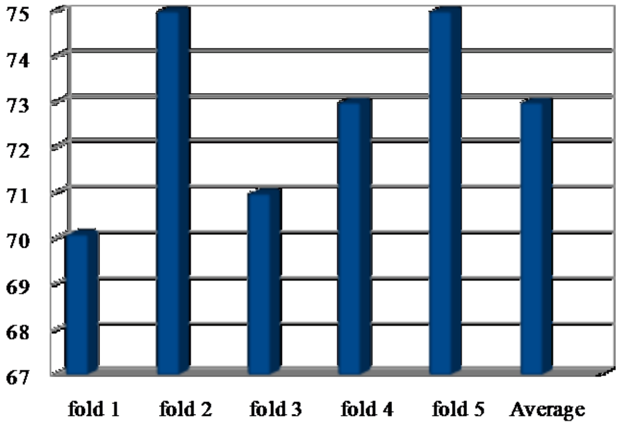

Figure 5. The result of dialog act classification model in 5 folds.

Table 1. The result of dialog act model in the best fold

Class	Human	Model	Match	Pre.(%)	Rec. (%)	F1-Score
inform	246	233	172	73.82	69.92	71.82
question	111	99	73	63.74	65.77	69.52
commis-sive	3	4	2	50.00	66.67	57.14
directive	53	65	37	56.92	69.81	62.71
none	231	243	172	73.82	69.92	71.82
Avg.1				65.05	69.32	67.12
Avg.2	644	644	456	70.81	70.81	**70.81**

Table 2 shows the result in a fold. The class column represents the classes. The human column is assigned labels of FS by humans. The model column shows the labels that the model predicted. The matching column is the number of predicted model labels matching the assigned label in the data. The following columns show the model's precision, recall and F1-score. Avg1 shows the accuracy of the model based on class. Avg2 shows the average accuracy of the labels based on examples.

II. Conclusion

In this research, we have researched dialog act recognition for Vietnamese conversational text. Our contributions can be summarized in the following points: (1) This is the first study of automatic dialog act recognition for Vietnamese conversational language; (2) We annotated dialog act according to ISO 24617-2; (3) We conducted an automatic recognition of conversational action with machine learning model Maximum entropy.

The results need to be improved with sophisticated features to achieve better results because of imbalance and noise data, and the inherent challenges of natural language understanding. On the other hand, in the future, we need to add more data and conduct a consensus assessment when labelling data.

Acknowledgment

This work was supported by the project T2019-07-07 from University of Information & Communication Technology (ICTU) – Thainguyen University.

References

[1] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. "*The design and implementation of xiaoice, an empathetic social chatbot*". arXiv preprint arXiv:1812.08989, 2018.

[2] Searle, John R. "*Austin on locutionary and illocutionary acts*." The philosophical review 77, no. 4, 1968, pp. 405-424.

[3] Susanne Jekat, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast, and J Joachim Quantz. "*Dialogue acts in verbmobil*", 1995.

[4] Stolcke A, Bratt H, Butzberger J, Franco H, Gadde VR, Plauché M, Richey C, Shriberg E, Sönmez K, Weng F, Zheng J. *"The SRI March 2000 Hub-5 conversational speech transcription system"*. In NIST Speech Transcription Workshop, 2000.

[5] Fang, Alex, J. Cao, H. Bunt, and Xiaoyue Liu. *"Applicability verification of a new ISO standard for dialogue act annotation with the Switchboard corpus."* In EACL Workshop on Innovative Hybrid Approaches to the Processing of Textual Data, 2012.

[6] Cervone, Alessandra, Giuliano Tortoreto, Stefano Mezza, Enrico Gambi, and Giuseppe Riccardi. *"Roving mind: a balancing act between open–domain and engaging dialogue systems."* Alexa Prize 1, 2017.

[7] Fang, Hao, Hao Cheng, Elizabeth Clark, Ariel Holtzman, Maarten Sap, Mari Ostendorf, Yejin Choi, and Noah A. Smith. *"Sounding board–university of washington's alexa prize submission."* Alexa prize proceedings, 2017.

[8] Yamamura, Takashi, Masato Hino, and Kazutaka Shimada. *"Dialogue act annotation and identification in a Japanese multi-party conversation corpus."* In Asia Pacific Corpus Linguistics, 2018.

[9] Yoshino, Koichiro, Hiroki Tanaka, Kyoshiro Sugiyama, Makoto Kondo, and Satoshi Nakamura. *"Japanese dialogue corpus of information navigation and attentive listening annotated with extended iso-24617-2 dialogue act tags."* In LREC, 2018.

[10] Kral, Pavel, Christophe Cerisara, Jana Kleckova, and Tomas Pavelka. *"Sentence structure for dialog act recognition in Czech."* In International Conference on Information & Communication Technologies, vol. 1, pp. 1214-1218. IEEE, 2006.

[11] Ji, Gang, and Jeff Bilmes. *"Backoff model training using partially observed data: Application to dialog act tagging."* In ACL, pp. 280-287, 2006.

[12] Kim, Hark-Soo, Choong-Nyoung Seon, and Jung-Yun Seo. *"Review of Korean speech act classification: machine learning methods."* Journal of Computing Science and Engineering 5, no. 4 (2011): 288-293.

[13] Elmadany, AbdelRahim A., and Walid Magdy Hamdy Mubarak. *"ArSAS: An Arabic speech-act and sentiment corpus of tweets."* In OSACT 3, pp. 20, 2018.

[14] Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard et al. *"The HCRC map task corpus." Language and speech* 34, no. 4 (1991): 351-366.

[15] Core, Mark G., and James Allen. *"Coding dialogs with the DAMSL annotation scheme."* In AAAI, vol. 56. 1997.

[16] John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. *"Switchboard: Telephone speech corpus for research and development."* In ICASSP-92, 1992.

[17] Jekat, Susanne, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast, and J. Joachim Quantz. *"Dialogue acts in VERBMOBIL."*, 1995.

[18] Bunt, Harry. *"Dynamic interpretation and dialogue theory."* The structure of multimodal dialogue 2, 1999.

[19] Bunt, Harry, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R. Traum. *"ISO 24617-2: A semantically-based standard for dialogue annotation."* In *LREC*, pp. 430-437. 2012.

[20] Liu, Yang, Kun Han, Zhao Tan, and Yun Lei. *"Using context information for dialog act classification in DNN framework."* In EMNL, 2017.

[21] Quarteroni, Silvia, Alexei V. Ivanov, and Giuseppe Riccardi. *"Simultaneous dialog act segmentation and classification from human-human spoken conversations."* In ICASSP, 2011.

[22] Quarteroni, Silvia, and Giuseppe Riccardi. *"Classifying dialog acts in human-human and human-machine spoken conversations."* In ISCA, 2010.

[23] Bunt, Harry, Volha Petukhova, Andrei Malchanau, Kars Wijnhoven, and Alex Fang. *"The dialogbank."* In LREC, 2016.

[24] Ngo, Thi-Lan, Pham Khac Linh, and Hideaki Takeda. *"A Vietnamese Dialog Act Corpus Based on ISO 24617-2 standard."* In LREC, 2018.

MSc Thi Lan Ngo is a PhD student at Vietnam Nationnal University, Hanoi and also a lecturer at Faculty of Information Technology, University of Information and Communication Technology, Thai nguyen City, Vietnam. Research interests: Computer science, Nature Language processing, Spoken Language Understanding.

MSc Thi Bich Ngoc Doan is a lecturer at Faculty of Information Technology, University of Information and Communication Technology, Thai nguyen City, Vietnam. Research interests: Computer science.

MSc Thi Lan Phuong Ngo is a lecturer at Faculty of Information Technology, University of Information and Communication Technology, Thai nguyen City, Vietnam. Research interests: Computer science.