2017 9th International Conference on Knowledge and Systems Engineering(KSE)

Dialogue Act Segmentation for Vietnamese Human-Human Conversational Texts

Thi–Lan Ngo
University of Information and Communication Technology
Thainguyen University (TNU)
Email: ntlan@ictu.edu.vn

Khac–Linh Pham
University of Engineering and Technology
Vietnam National University, Hanoi (VNU)
Email: phamkhaclinh2017@gmail.com

Minh–Son Cao
University of Engineering and Technology
Vietnam National University, Hanoi (VNU)
Email: soncm_58@vnu.edu.vn

Son–Bao Pham
University of Engineering and Technology
Vietnam National University, Hanoi (VNU)
Email: sonpb@vnu.edu.vn

Xuan–Hieu Phan
University of Engineering and Technology
Vietnam National University, Hanoi (VNU)
Email: hieupx@vnu.edu.vn

***Abstract*—Dialog act identification plays an important role in understanding conversations. It has been widely applied in many fields such as dialogue systems, automatic machine translation, automatic speech recognition, and especially useful in systems with human-computer natural language dialogue interfaces such as virtual assistants and chatbots. The first step of identifying dialog act is identifying the boundary of the dialog act in utterances. In this paper, we focus on segmenting the utterance according to the dialog act boundaries, i.e. functional segments identification, for Vietnamese utterances. We investigate carefully functional segment identification in two approaches: (1) machine learning approach using maximum entropy (ME) and conditional random fields (CRFs); (2) deep learning approach using bidirectional Long Short-Term Memory (LSTM) with a CRF layer (Bi-LSTM-CRF) on two different conversational datasets: (1) Facebook messages (Message data); (2) transcription from phone conversations (Phone data). To the best of our knowledge, this is the first work that applies deep learning based approach to dialog act segmentation. As the results show, deep learning approach performs appreciably better as to compare with traditional machine learning approaches. Moreover, it is also the first study that tackles dialog act and functional segment identification for Vietnamese.**

Keywords*—*Dialog act segmentation, functional segment, Vietnamese conversation.

I. Introduction

Automatic recognition of user intent from utterances in their interaction with systems through the conversational interface is a very challenging task that has attracted a lot of attention from research community for two decades. The goal is to design methods to make computers interact more naturally with human beings. Identifying dialog acts (DAs) within an utterance, i.e. identifying its illocutionary act of communication, plays a key role in understanding user's intent. Because, *"Dialog act is a communicative activity of dialog participant, interpreted as having a certain communicative function and semantic content"* [1]. It presents meaning of utterances at the discourse level. It is a complementary process to concept extraction. Therefore, it is essential for the complete understanding of conversations. It is important for many applications: dialogue systems, automatic translation machine [2], automatic speech recognition, etc [3] [4] and has been studied in various languages such as English, Chinese, Arabic, Czech, Korean. Whilst in Vietnamese languages, dialog act has only been studied in linguistics, our work in this paper is a preliminary study about automatic identification of dialog act, as well as dialog act segmentation.

Prior to DA identification, utterances must be segmented according to DA boundaries. In the past, there have been studies of DA segmentation such as Umit Guz et al. implemented DA segmentation of speech using multi-view semi-supervised learning [5]; Jeremy Ang et al. explored DA segmentation using simple lexical and prosodic knowledge sources [6]; Warnke et al. calculated hypotheses for the probabilities exceeded a predefined threshold level in VERBMOBIL corpus [7]; Silvia Quarteroni et al. segmented human-human dialog into turns and intra-turn segmentation into DA boundaries using CRFs to learn models for simultaneous segmentation of DAs from whole human-human spoken dialogs [8]. These studies segmented turns into sentence unit to do dialog act segmentation. In my work, different from those studies, we segment utterances into the smallest meaningful units – *"functional segment"* unit. According to ISO 24617-2 standard about Dialog Act, a functional segment (FS) is defined as "minimal stretch of communicative behavior that have a communicative function" [1]. For example, in the utterance "xin chào cậu khỏe chứ" ("hello are you fine"), there are two functional segments: "xin chào" ("hello") (its dialog act is greeting), and "cậu khoẻ chứ" ("are you fine") (its dialog act is check question). We investigate thoroughly functional segment identification in two approaches: (1) machine learning approach with ME, CRF; (2) deep learning approach with Bi–LSTM–CRF. Recently, ME, CRF and Bi–LSTM–CRF have been applied to a variety of sequence labeling and segmentation tasks in Natural Language Processing and have achieved state-of-the-art results [9]. Therefore, we expect that these methods apply to the FS identification task for Vietnamese can make similar successes. To do the task, we first build two annotated corpus from Facebook messages and transcription from phone conversations. For a careful evaluation, different ME, CRF and Bi–LSTM–CRF models were trained and their results are compared and shown contrast with each other. Moreover, we also show the characteristics of two different conversational data sets and their effect on the experimental results of the task of the dialog act segmentation task.

We can summary our main contributions in this paper in two aspects:

- First, we built two Vietnamese conversational text datasets which are segmented into FSs based on FS concept from the ISO standard and ready to contribute to the DialogBank [1] for Vietnamese. We also built online chat dictionary which contains abbreviations, slang words and teen code and Vietnamese local dialect dictionary.
- Second, two machine learning techniques and a deep learning technique are applied and compared on the task of automatic dialog act segmentation. Deep learning technique is also applied for the first time to dialog act segmentation. The results of the deep learning technique are very promising, opening up a new way to approach dialog act segmentation and dialog act in general for applications for future studies.

The rest of the paper is organized as follows: Section II presents briefly background about FS formation in Vietnamese conversational texts and units of a dialogue. In Section III we describe our two human-human conversation corpus. We also discuss the impact of our conversational data sets to the functional segment identification task in this section. We describe quickly the two learning models ME, CRF and the deep learning model, Bi–LSTM–CRF for labeling and segmenting FS in Section IV. Section V mainly presents the framework of using MEs, CRFs, Bi–LSTM–CRF for Vietnamese FS segmentation and result comparison and evaluation. Finally, Section VI shows some conclusions and the work that need research in the future.

II. Backgroud: Functional segment and units of a dialogue

DAs are extended from the speech act theory of Austin [10] and Searle [11] to model the conversational functions that utterances can perform. It is the meaning of an utterance at the level of illocutionary force, such as statement, question and greeting. Detection of dialog acts need to perform: 1) the segmentation of human–human dialogues into turns, 2) the intra-turn segmentation into DA boundaries, i.e. functional segment identification and 3) the classification of each segment according to a DA tag [12].

In which, "turn", "dialog act", "functional segment" terms are defined slightly different between different domains and different purposes. But these are standardized and united in ISO standards as follows:

Turn:

A "turn" is definite as *"stretch of communicative activity produced by one participant who occupies the speaker role bounded by periods where another participant occupies the speaker role"*. Dialogue participants (sender, addressee) normally take turns in conversation. Several utterances from one of the dialogues in our corpus are shown as examples of *Turn*, *Message*, and *Functional segment* in Table I and Table II.
In our Message data, a turn is seen as a collection of continuous messages sent by one participant. In which, a message is defined as a group of words that are sent from one dialogue participant to the other. For instance, turn t_2 includes four messages ms_2, ms_3, ms_4, ms_5 (Table I).

Functional segment:

A functional segment is the *"minimal stretch of communicative behavior that has a communicative function"*, *"minimal in the sense of not including material that does not contribute to the expression of the function or the semantic content of the dialogue act"* [1]. A functional segment may be shorter than turns and continuous, for example as in Table I, t_1 includes two functional segments fs_1 and fs_2. A functional segment may be discontinuous, with examples such as fs_4 and fs_{10}. fs_5 is nested within fs_4. In addition, functional segment fs_{10} is combined from two messages, fs_8 overlaps fs_{10}. Thus, we can see that a functional segment may be continuous, may be discontinuous, may be overlapped and nested. The detailed explanation of the types of FS is presented in [13] and the ISO 24617-2 standard.

Dialog Act:

DA is *"communicative activity of a dialogue participant, interpreted as having a certain communicative function and semantic content"*. For example:

"xin chào cậu khoẻ chứ" ("hello are you fine")

DAs of "xin chào" (hello) are *Greeting* and *Opening*. DA of "cậu khoẻ chứ" ("are you fine") is *Check Question.*

III. Corpus Building: Message data & Phone data

In Vietnamese, there is no publicly available standard corpus. Therefore we need to build first a reference corpus for training and evaluation. For this work, we have to build two corpora of data from human-human conversations in various domains. One is chat texts and other is spoken texts.

A. Message corpus

Our Message data set is collected from Facebook messages of 20 volunteers. The data set contains 280 human-human Vietnamese dialogues in any topics with a total number of 4583 messages. The average length of dialogues is 16.4 messages. The data set was independently labeled by three annotators. The agreement score of our data set achieved 0.87 Fleiss' kappa measure [14]. As observed from our data, there are some challenges as follows:

1) The data is very noisy because it contains many acronyms, misspellings, slang, and emoticons. These

[1] https://dialogbank.uvt.nl/

Table I. EXAMPLES OF FUNCTIONAL SEGMENT AND TURN IN MESSAGE DATA.

Participants	Messages	Turns	Functional segments	Type
S	Đây là đề tài chung tôi sẽ hưỡng dẫn bạn dần dần :) (This is the general topic I will guide you gradually :)) (ms1)	Đây là đề tài chung tôi sẽ hướng dẫn bạn dần dần :) (This is the general topic I will guide you gradually :)) (t1)	Đây là đề tài chung (This is the general topic) (fs1)	continuous
			Tôi sẽ hướng dẫn bạn dần dần :) (I will guide you gradually :)) (fs2)	continuous
A	uhhhhhh nhưng thời gian (Yessssss, but the time) (ms2)	uhhhhhh nhưng thời gian hic hic ngắn quá sợ k làm đc (Yessssss, but the time is too short I am afraid of can not done) (t2)	uhhhhhh (fs3)	continuous
A	hic hic (ms3)		nhưng thời gian ngắn quá (but the time is too short) (fs4)	discontinuous
A	ngắn quá (too short) (ms4)		hic hic (fs5)	nested
A	sợ k làm đc (I am afraid of can not done) (ms5)		sợ k làm đc (I am afraid of can not done) (fs6)	continuous
S	Cậu còn chưa bắt đầu mà đã sợ rồi (You have not started yet, have you been afraid) (ms6)	Cậu còn chưa bắt đầu mà đã sợ rồi (You have not started yet have been afraid) (t3)	Cậu còn chưa bắt đầu mà đã sợ rồi (You have not started yet have been afraid) (fs7)	continuous
A	chưa bắt đầu hic :3 cái gì? (not started yet hic :3 what?)(ms7)	chưa bắt đầu hic :3 cái gì ? tôi đang làm rồi mà (not started yet hic :3 what? I am doing) (t4)	chưa bắt đầu (fs8)	overlap
A	Tôi đg làm rồi mà :) (ms8)		hic :3 (fs9)	continuous
			Chưa bắt đầu cái gì? (not started yet hic :3 what?) (fs10)	overlap and discontinuous
			Tôi đg làm rồi mà :) (I am doing) (fs11)	continuous

Table II. EXAMPLES OF FUNCTIONAL SEGMENT AND TURN IN PHONE DATA.

Participants	Turn	Functional segment	Type
S	ở <no speech> ở quê có những đặc sản gì vậy anh (in <no speech> in home town What is the specialty) (t1)	ở (in) (fs1)	continuous
		ở quê có những đặc sản gì vậy anh (in home town What is the specialty) (fs2)	continuous
A	cái này a - ủa cái này thì củng nói thật chứ nhiều đặc sản lắm <no speech> đặc sản quê hương là mỗi nơi mỗi khác <no speech> (About this I – oh about this then being honest there are a lot specialties specialties of each country is different) (t2)	cái này a - (this) (fs3)	continuous
		ủa cái này thì củng nói thật chứ nhiều đặc sản lắm (About this I – oh about this then being honest there are a lot specialties) (fs4)	continuous
		đặc sản quê hương là mỗi nơi mỗi khác (specialties of each country is different) (fs5)	continuous
S	dạ vâng ạ (yes yes) (t3)<no speech>	dạ (yes) (fs6)	continuous
		vâng ạ (yes) (fs7)	continuous
A	ở trên này thì có là là nói chung là như là ở sông thì có cá sông nả <no speech> cá sông là tuyệt vời nhức hes chơ mà dưới biển thì có cá biển <laugh> nhưng mà ở sông thì lại lại lại chuộng cấy cá (Over here there are are in general there are are like river has river fishes <nospeech>river fishes are the best but in sea we also have sea fishes but near river also also also prefer sea fishes <laugh>) (t4)	ở trên này thì có như là ở sông thì có cá sông nả (Over here there are are in general there are are like river has river fishes) (fs8)	discontinuous
		là là nói chung là (is is in general) (fs9)	nested
		cá sông là tuyệt vời nhức (river fishes are the best) (fs10)	continuous
		chơ mà dưới biển thì có cá biển (but in sea we also have sea fishes)(fs11)	continuous
		nhưng mà ở sông thì lại chuộng cấy cá (but near river also also also prefer sea fishes) (fs12)	continuous
		lại lại (fs13)	nested

informal natures of chat text, which make conventional features such as punctuation mark, part–of–speech (POS), syntax of sentence and capitalization, are not reliable. Text message conversations are often written with non-standard word spellings. While some of them are unintentional misspellings, many of them are purposely produced, for example,
S: "đi chơi điiiiiii" ("let's go outttttt")
A:"không đang ốm quá !!!!!!!! (" no I am too sick !!!!!!").
The intent of the utterance by person S that: he want to express more clearly his desire by using non-standard form "iiiiiiii" instead of the standard "i". If the non-standard form was normalized to the standard form, in this case, the intent conveyed by the utterance would be ambiguous; "iiiiiii" could suggest that person S is very excited to go out with person A. The non–standard word forms that contain additional pragmatic information presented in the non-standard form should be retained in the data pre–processing stage.

2) The message's short nature leading to the availability of very limited context information.
3) In text chat dialogue, end of a turn is not always obvious. A turn often contains multiple messages. A message is often in a clause or utterance boundaries, but it is not always correct. Therefore, although the boundary of a message can be a useful feature to

FS identification but sometimes a FS may contain multiple messages, and even may include only a part of one message and a part of the next message. This indistinct end of a turn also leads to the end of a misleading message. In sudden interruption cases, messages can become out of sync. Each participant tends to respond to a message earlier than the previous one, making the conversation also being out of order and the conversation seem inconsistent when read in sequence. This is a difficult problem for processing the dialog act segmentation.

In short, unlike carefully authored news text, conversational text poses a number of new challenges, due to their short, context-dependent, noisy and dynamic nature. Tackling this challenge, ideally, requires changing related natural language processing tools to become suitable for texts from social media network or normalizing conversational texts to fit with existing tools. However, both of which are hard tasks. In the scope of this paper, we standardize the message data using our online chat dictionary to match popular abbreviations, acronyms, and slang with standard words in the pre-processing stage.

Online chat dictionary

Our online chat dictionary includes abbreviations, slang and the words that are written in teen style (teen code) such as “bj"- “bây giờ" (“now"), “ck" - “chồng" (“husband"), “4u" - “cho bạn" (“for you"). The letters “c", “k", “q" are usually replaced by “k", “ch” but often replaced with “ck” ... Using online chat dictionary to standardize the message data, the noisiness of input data will be reduced. This make it more formal and help the models run better.

B. Phone corpus

Our Phone data set is build from scripted telephone speech of LDC2017S01 data (IARPA Babel Vietnamese Language Pack IARPA-babel107b-v0.7 [2]). LDC2017S01 contains Vietnamese phone audios and transcripts. The Vietnamese conversations in these corpus contain different dialects that spoken in the North, North-Central, Central and Southern regions in Vietnam. We selected 22 conversations and segment its transcripts into the turn by manual. Then, the turns are annotated FS. The Phone data includes 1545 turns and 3500 FSs with an average of 70 turns and 160 FSs per conversation. The agreement scores of the phone data set is 0.84 Fleiss' kappa measure.

FS recognition for spoken texts, however, is more challenging than working with written documents due to some reasons as follows:

1) First, spoken text are commonly shorter and less grammatical, not comply with rigid syntactic constraints. Sentence elements like subject or object are often omitted. It is very context-dependent. Also, there are no punctuation marks in the texts. It, therefore, is non–trivial to segment and parse spoken sentences correctly.
2) Second, conversational speech contains a lot of self-correcting, hesitation, and stutter. This is one of the main reasons that causes nested FS. fs_9 and fs_{13} within turn t_4 in Table II are the instances.
3) Third, the output text of Automatic Speech Recognition are all in lowercase and bearing a small percentage of errors.

These challenges make it extremely difficult to recognize FS in particular and in understanding spoken language in general.

Vietnamese local dialect dictionary

The LDC2017S01 data is built from spoken conversations in the North, North-Central, Central and Southern Vietnamese dialect. Because of the nature of Vietnamese dialects, a lot of words in local dialects can be changed to standard dialect (the North Vietnamese dialect) without affecting the meaning of the utterances in which they belongs. For instances, “Răng rứa" means “sao thế" (what up); “Mi đi mô" means “Mày đi đâu" (where are you going?). Therefore we created a dictionary to match these words with standardized words. By doing so, the data sets become more uniform. This makes it easier to handle and help the models to run better. Our dictionary is not only useful in this study but also can be very helpful in all other studies that involve Vietnamese human–human, and human–machine conversation.

IV. DA SEGMETATION WITH ME, CRF AND BI-LSTM-CRF

The number of discontinuous or nested functional segments account for a very small percent in both data sets (0.5% in the Message corpus, 0.9% in the Phone corpus). Hence there are not enough discontinuous or nested functional segments so the models can learn to identify them. For that reason, this paper only focuses on identifying continuous and unnested functional segments (which make up more than 99% of both data sets). In future studies, we intend to increase the size of our data sets, the number of discontinuous or nested functional segments and study methods to identify these functional segments. In this paper, we cast the segmentation problem as a sequential tagging task: the first word of a FS is marked with B_fs (Begin of a FS), the token that is inside of a FS is marked with I_FS (Inside of a FS). The problem of FS identification in a sentence is modeled as the problem of labeling syllables in that sentence with two above labels.

Let $t = \{t_1, t_2, ...t_n\}$ be turns and $y = \{B, I\}$ be per-token output tags. We predict the most likely y, given a conditional model $P(y|t)$.

A. Maximum Entropy

The ME (Maxent) model defines conditional distribution of class (y) given an observation vector t as the exponential form in Formula (1) [15]:

$$P(y/t) = \frac{1}{Z(t)} exp\left(\sum_{1}^{K} \theta_k(t, y)\right) \quad (1)$$

where θ_k is a weight parameter to be estimated for the corresponding feature function $f_k(t, y)$, and $Z(t)$ is a normalizing factor over all classes to ensure a proper probability. K is the total number of feature functions. We decided to use ME for evaluation and comparison because it is commented that it is suitable for sparse data like natural language, encode various rich and overlapping features at different levels of granularity [16].

[2] https://catalog.ldc.upenn.edu/LDC2017S01

B. Conditional Random Fields

The CRFs model defines also the conditional distribution of the class (y) given an observation vector t as the Formular (1) [17]. In which θ_k is a weight parameter to be estimated for the corresponding feature function $f_k(t, y)$, and $Z(t)$ is a normalizing factor over all classes to ensure a proper probability. And K is the total number of feature functions. It is essentially a ME model over the entire sequence. It is unlike the Maxent above since it models the sequence information, because the Maxent model decides for each state independently with the other states. For example, a transcription utterance together with class tags used for the CRF word detection model in Dialog act segmentation as follows:

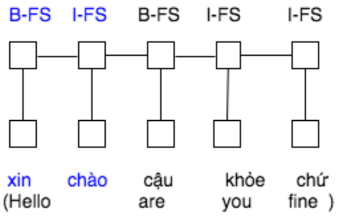

Figure 1. A CRF model for identifying FS.

Training ME and CRF are commonly performed by maximizing the likelihood function with respect to the training data using advanced convex optimization techniques like L–BFGS [18].

C. Deep learning–based models with Bi–LSTM–CRF

Bi–LSTM–CRF network is formed by combining a bidirectional LSTM network and a CRF network [9]. Therefore Bi–LSTM–CRF can efficiently use past and future input features via a Bi–LSTM layer and sentence level tag information via a CRF layer. A CRF layer is represented by lines which connect consecutive output layers. A CRF layer has a state transition matrix as parameters. The following are examples of a text in the Bi–LSTM–CRF model: BI–LSTM–CRF has emerged as a standard method for obtaining per-token vector representations serving as input to various token labeling tasks. We expect that dialog act segmentation in Vietnamese using BI–LSTM–CRFs model will also similar to highly accurate results.

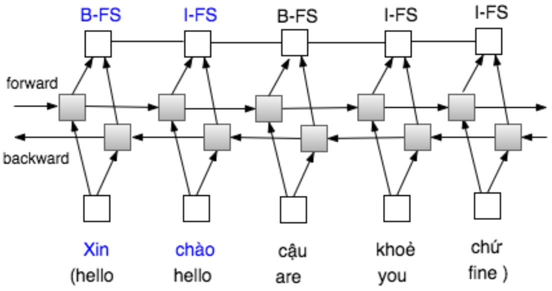

Figure 2. A BI-LSTM-CRF model for identifying FS.

V. Evaluation

The simple lexical feature, n–gram (unigram, bigram and trigram), is used for the ME and CRF models. We do experiments on two different conversational data sets (Message data set and Phone data set) after normalizing these data sets using local dialect dictionary and online chat dictionary.

Training ME and CRF are commonly performed by maximizing the likelihood function with respect to the training data using quasi-Newton methods like L–BFGS [18]. Thus, in the experiments with ME and CRF, we use L-BFGS method. For CRF models, we use second-order Markov dependency. On experiment with CRF, we use tools: FlexCRFs - a C/C++ implementation of CRFs [3]. On experiment with Bi–LSTM–CRF, our setup is based on study of Lample et al. [4] [19] .

For evaluating each experiments, we randomly divide each corpus into five parts to do 5-fold cross-validation test. In each fold we take one partition for testing and 4 partitions for training. The summary of the experiment results on Message data set is shown in Table III, the experiment results on Phone data set is shown in Table IV.
The results of label-based performance evaluation are significantly higher than the results of label-based performance evaluation and chunk-based performance evaluation. The evaluation measures for this task are precision and recall based on labels:

$$precision = \frac{number\ of\ correctly\ predicted\ label\ by\ the\ model}{number\ of\ label\ predicted\ by\ the\ model};$$

$$recall = \frac{number\ of\ correctly\ predicted\ label\ by\ the\ model}{number\ of\ actual\ label\ annotated\ by\ humans};$$

$Average_{macro}$ is the average of the precision and recall of the model on different classes. $Average_{micro}$ is sum up the individual true positives, false positives, and false negatives of the model for different classes.

The precision and recall based on chunks is as follows:

$$precision = \frac{number\ of\ correctly\ predicted\ FS\ by\ the\ model}{number\ of\ FS\ predicted\ by\ the\ model};$$

$$recall = \frac{number\ of\ correctly\ predicted\ FS\ by\ the\ model}{number\ of\ actual\ FS\ annotated\ by\ humans};$$

F_1- score in the both of evaluations is calculated as follows:

$$F_1 = \frac{2*(precision*recall)}{(precision+recall)};$$

BI–LSTM–CRF models achieved the highest performance (average F1 of 90.42% with Messages dataset, 73.26% with Phone dataset). This was an indication that it is robust and less affected by the removal of engineering features.

Performance results with Messages data (manual texts) are higher than results achieved with Phones data (Automatic Speech Recognition transcripts) because turns in Messages data set are often shorter and less ambiguous for dialog act segmentation than turns in Phone data set. Turns in Phone data set also includes hesitance, repeat, and overlap. These make discontinuous segments, either within a turn or spread over several turns as we have already discussed. A greater challenge is posed by those cases where different functional segments overlapped.

Another observation from the results is that Bi-LSTM-CRFs, the deep learning approach, performs significantly bet-

[3] http://flexcrfs.sourceforge.net/

[4] https://github.com/glample/tagger

Table III. PERFORMANCE COMPARISON AMONG ME, CRF AND BI–LSTM–CRF MODELS ON MESSAGE DATASET.

Model	Lable	Precision	Recall	F1-score
ME	B-fs	77.36	77.74	77.54
	I-fs	94.67	94.56	94.61
	$Average_{macro}$	86.01	86.15	**86.08**
	$Average_{micro}$	91.31	91.31	**91.31**
	Chunk	57.38	57.33	**57.34**
CRF	B-fs	100	80.03	88.9
	I-fs	95.46	100	97.68
	$Average_{macro}$	97.73	90.01	**93.71**
	$Average_{micro}$	96.16	96.16	**96.16**
	Chunk	83.8	67.08	**74.51**
BI-LSTM-CRF	B-fs	97.11	95.24	96.17
	I-fs	98.87	99.32	99.1
	$Average_{macro}$	97.99	97.28	**97.64**
	$Average_{micro}$	98.54	98.54	**98.54**
	Chunk	91.3	89.56	**90.42**

Table IV. PERFORMANCE COMPARISON AMONG ME, CRF AND BI–LSTM–CRF MODELS ON PHONE DATASET.

Model	Lable	Precision	Recall	F1-score
ME	B-fs	83.51	75.89	79.52
	I-fs	93.9	96.12	95
	$Average_{macro}$	88.7	86.01	**87.34**
	$Average_{micro}$	91.96	91.96	**91.96**
	Chunk	61.88	56.23	**58.92**
CRF	B-fs	95.22	71.24	81.43
	I-fs	93.09	99.06	95.98
	$Average_{macro}$	94.15	85.15	**89.42**
	$Average_{micro}$	93.4	93.34	**93.37**
	Chunk	66.82	50.18	**57.27**
BI-LSTM-CRF	B-fs	94.38	84.6	89.22
	I-fs	96.02	98.64	97.31
	$Average_{macro}$	95.2	91.62	**93.38**
	$Average_{micro}$	95.7	95.7	**95.7**
	Chunk	77.48	69.47	**73.26**

ter than both CRF and ME, the machine learning approaches, by every measure. Because deep learning has never been used for dialog act segmentation before, this result opens up a very promising new direction for future studies to approach dialog act segmentation and dialog act in general. Between the machine learning approaches, CRF performs better than ME overall. This can be explained by looking at how CRF and ME works. ME is locally re-normalized and suffers from the label bias problem, while CRFs are globally re-normalized. This label bias problem can happen a lot, especially with very context-dependent data sets like Message corpus and Phone corpus.

VI. CONCLUSIONS

We have presented a thorough investigation on Vietnamese FS identification using machine learning approach and deep learning approach. We built two annotated corpora for evaluation and two dictionaries that make the data sets more uniform and help the models run better. Two machine learning techniques and a deep learning technique are applied and compared on the task of automatic dialog act segmentation. Deep learning technique is also applied for the first time to dialog act segmentation. We also draw some useful conclusions observed from the experimental results that can be very helpful for future studies.

These encouraging results show that the task of identifying functional segment is promising to continue to the next dialogue act identification steps and towards understanding intentions in the users' utterances for Vietnamese. For future work, we intend to extend the studies into two directions. First, we plan to increase the size of our data set to get sufficient amount of instances in different types of functional segment and study deeper methods to solve nested FS identification. Second, we intend to use features included in the data sets as dialogue history, prosody to improve automatic FSs recognition and dialogue processing.

ACKNOWLEDGMENT

This work was supported by the project QG.15.29 from Vietnam National University, Hanoi (VNU).

REFERENCE

[1] Bunt, H., Alexandersson, J., Choe, J.W., Fang, A.C., Hasida, K., Petukhova, V., Popescu-Belis, A., Traum, D., "ISO 24617- 2: A Semantically-Based Standard for Dialogue Annotation." In: LREC'12, 2012.

[2] Tanaka, H., Yokoo, A. "An efficient statistical speech act type tagging system for speech translation systems." In ACL'37, 1999.

[3] Y. Wang, L. Deng, and A. Acero, "Spoken language understanding: An introduction to the statistical framework." IEEE Signal Processing Magazine, vol. 22, no. 5, pp. 16–31, 2005.

[4] Král, P., Cerisara, C. "Dialogue act recognition approaches." In Computing and Informatics, 2012.

[5] Guz, U., Cuendet, S., Hakkani-Tur, D., Tur, G. "Multi-view semi-supervised learning for dialog act segmentation of speech." IEEE transactions on audio, speech, and language processing, 18(2), 320-329, 2010.

[6] Ang, Jeremy, Yang Liu,Elizabeth Shriberg."Automatic dialog act segmentation and classification in multiparty meetings." In: ICASSP'05. IEEE International Conference on. Vol, 2005.

[7] Warnke, V., Kompe, R., Niemann, H., Nöth, E. "Integrated dialog act segmentation and classification using prosodic features and language models." In: Eurospeech, 2997.

[8] Quarteroni, S., Ivanov, A. V., Riccardi, G. "Simultaneous dialog act segmentation and classification from human-human spoken conversations." In: ICASSP, pp. 5596-5599, 2011.

[9] Huang, Z., Xu, W., Yu, K. "Bidirectional LSTM-CRF models for sequence tagging." In: arXiv preprint arXiv:1508.01991, 2015.

[10] Austin, J. L. "How to do things with words." In: Oxford university press, 1975.

[11] Searle, J. R. "A taxonomy of illocutionary acts." 1975.

[12] Ramacandran, Nithin. "Dialogue Act Detection from Human-Human Spoken Conversations." In: International Journal of Computer Applications, 2013.

[13] Bunt, H., "Multifunctionality in dialogue." In: Computer Speech & Language 25.2, 222-245, 2011.

[14] Fleiss, J. L. "Measuring nominal scale agreement among many raters". In: Psychological bulletin, 76(5), 378, 1971.

[15] Berger, A., Pietra, S.A.D., Pietra, V.J.D. "A maximum entropy approach to natural language processing." In: Computational Linguistics, 22(1), 39–71, 1996.

[16] Nigam, K., Lafferty, J., McCallum, A.: "Using maximum entropy for text classifi- cation. In: IJCAI Workshop on Machine Learn. for Info. Filtering, pp.61–67, 1999.

[17] Lafferty, J.D., McCallum, A., Pereira, F. "Conditional random fields: probabilistic models for segmenting and labeling sequence data." In: ICML, pp.282–289, 2001.

[18] Liu, D., Nocedal, J.: "On the limited memory BFGS method for large–scale opti- mization." In: Mathematical Programming, 45, pp.503–528, 1989.

[19] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C. "Neural architectures for named entity recognition." In: arXiv preprint arXiv:1603.01360, 2016.