Arabic Dialogue Act Recognition for Textual Chatbot Systems

Alaa Joukhadar
Higher Institute for Applied Sciences and Technology
Damascus, Syria
alaa.joukhadar@hiast.edu.sy

Huda Saghergy, Leen Kweider
IT Engineering Faculty,
Damascus University
huda.saghergy@gmail.com
leenkweider@gmail.com

Nada Ghneim
Higher Institute for Applied Sciences and Technology,
Damascus, Syria
nada.ghneim@hiast.edu.sy

Abstract

Automatic Dialogue Acts Recognition is considered a crucial step for semantic extraction in Natural Language Understanding and Dialogue Systems. In this paper, we introduce our work aiming to recognize the dialogue acts of the users in a Textual Dialogue system using Levantine Arabic dialect. Our Dialogue acts have 8 types: Greeting, Goodbye, Thanks, Confirm, Negate, Ask_repeat, Ask_for_alt, and Apology. Various Machine Learning algorithms -with different features have been used to detect the correct speech act categories: Logistic Regression, SVM, Multinomial NB, Extra Trees Classifier, Random Forest Classifier. We also used the Voting Ensemble method to make the best prediction from each classifier. We compared the results of the proposed models on a hand-crafted corpus in the restaurants orders and airline ticketing domain. The SVM algorithm with 2-gram has given the best results.

1 Introduction

Modeling and automatically identifying the structure of spontaneous dialogues is very important to better interpret and understand them. Speech act (or Dialogue act) recognition is considered an essential step in these models. Austin defines in (Austin, 1962) the speech act as the meaning of an utterance at the level of illocutionary force. In other words, the dialogue act is the function of a sentence (or its part) in the dialogue. For example, the function of a question is to request some information, while an answer shall provide this information.

The recognition of speech acts has gained considerable interest over the past two decades. Its significance derives from its broad range of applications such as: Tutorial Dialogue Systems (Ezen-Can and Boyer, 2014) (Rus et al., 2017), Machine Translation (Fukada et al., 1998), Animation of Talking Heads, Conversational Analysis (Fišel, 2007), Natural Speech Synthesis, Customer Service Conversation Outcomes Prediction (Oraby et al., 2017), etc.

Many researchers have proposed different approaches to recognize speech acts in different languages, such as English (Bothe et al., 2018) (Chen et al., 2018) (Elmadany et al., 2018), Korean (Kim et al., 2011) (Kim and Kim, 2018), German (Zarisheva and Scheffler, 2015), etc. They have developed different tag sets and corpora, investigating a variety of supervised (Tavafi et al., 2013) (Chen et al., 2018) (Kumar et al., 2018) and unsupervised machine learning techniques (Ezen-Can and Boyer, 2014) (Kristy Elizabeth Boyer, 2015) (Sherkawi et al., 2018).

The correct interpretation of the intents behind a speakers utterances plays a very essential role in determining the success of a dialogue. Therefore, the intents classification module lies at the very core of any dialogue system.

In general, chat bot systems can be composed of three basic components: Natural Language Understanding (NLU), Dialogue Manager (DM), and Natural Language Generation (NLG). The recognized dialogue acts (from the Natural Language Understanding component) are usually used as an input to the Dialogue Manager component, to help determine the next action of the system, such as giving correct information when the user is asking a question, and keeping quiet when the user is just acknowledging, or giving a simple comment. The Dialogue Acts taxonomy differs according to the

dialogue system domain.

The work presented here is part of a project that aims to build a domain-independent textual dialogue system in Levantine Arabic dialect. The concept of dialect in Arabic world is different from what is known in the west, as people do not use Standard Arabic in their day life but different dialects, which are very different from standard Arabic. Arabic dialects are generally classified by regions, such as in (Habash, 2010) where Arabic dialects were classified into North African, Levantine, and Egyptian. Our work considers the dialogues in Levantine (mostly Syrian) dialect.

The main contributions of this work are as follows:

- We provide an insight on the annotation of our Levantine Arabic Dialogue Act corpora used in restaurants orders and airline ticketing domain.
- We propose 5 learning models for Dialogue Act identification, along with different features.
- We evaluate and compare the accuracy of the different models on our Dialogue Act dataset.

Our paper is divided as follows: section 2 presents related works, section 3 is our proposed methods for the classification of the speech acts, including the proposed taxonomy and dataset. Section 4 presents the evaluation for our approach, and section 5 is the conclusion.

2 Related Works

Automatic recognition of dialogue acts is an important, yet still underestimated component of Human-Machine Interaction dialogue architecture. The research in this area have made great progress during the few last years.

In (Kumar et al., 2018) authors have built a hierarchical recurrent neural network using bidirectional LSTM as a base unit and the conditional random field (CRF) as the top layer to classify each utterance into its corresponding dialogue act. The hierarchical network learns representations at word, utterance, and conversation levels. The conversation level representations are input to the CRF layer, which takes into account all previous utterances and their dialogue acts. They validated their approach on Switchboard (SwDA) and Meeting Recorder Dialogue Act (MRDA) data sets, and show performance improvement over the state-of-the-art methods by 2.2% and 4.1% absolute points, respectively.

(Bothe et al., 2018) used simple RNN to model the context of preceding utterances. They used the domain independent pre-trained character language model to represent the utterances. Their proposed model was evaluated on the Switchboard Dialogue Act corpus and achieved an accuracy of 77.34% with context compared to 73.96% without context.

(Lee et al., 2016) have also presented a model based on recurrent neural networks and convolutional neural networks that incorporates the preceding short texts. They validated their model which achieved state-of-the-art results on three different datasets (DSTC 4, MRDA, and SwDA) for dialogue act prediction.

(Khanpour et al., 2016) have applied a deep LSTM structure to classify dialogue acts (DAs) in open-domain conversations (Khanpour et al., 2016). They found that the word embeddings parameters, dropout regularization, decay rate and number of layers have the greatest impact on the final system accuracy. They validated their model which outperformed the state-of-the-art on the Switchboard corpus by 3.11%, and MRDA by 2.2%.

In (Chen et al., 2018) authors proposed the CRF-Attentive Structured Network (CRF-ASN) to solve the problem in two steps. They first encoded the rich semantic representation on the utterance level by incorporating hierarchical granularity and memory enhanced inference mechanism. The learned utterance representation captured long term dependencies across the conversation. Next, they adopted the internal structured attention network to compute the dialogue act influence and specify structural dependencies in a soft manner. The approach enabled the soft-selection attention on the structural CRF dependencies and took account of the contextual influence on the nearing utterances. The method achieved better performance than several state-of-the-art solutions on SwDA and MRDA datasets.

(Wan et al., 2018) proposed an improved dynamic memory networks with hierarchical pyramidal utterance encoder. Moreover, they applied adversarial training to train the proposed model, which was evaluated on Switchboard dialogue act corpus and the MapTask corpus. Extensive ex-

periments showed that the model was robust and achieved better performance compared with some state-of-the-art baselines.

Concerning non English languages, some researches were focused on multilingual domain, such as the work of (Cerisara et al., 2018) who proposed a deep neural network approach that explores recurrent models to capture word sequences within sentences, and further studied the impact of pre-trained word embeddings. The model was validated on three languages: English, French and Czech, and the performance was consistent across these languages and comparable to the state-of-the-art results in English.

(Jahanbakhsh-Nagadeh et al., 2019) presented a dictionary-based statistical technique for Persian speech acts recognition. They used lexical, syntactic, semantic, and surface features to detect seven classes of speech acts. To evaluate their proposed technique, they implemented four classification methods including Random Forest, Support Vector Machine, Naive Bayes, and K-Nearest Neighbors. The experimental results demonstrated that the proposed method using RF and SVM had the best classification accuracy.

Arabic speech acts classification started to show few initiatives. Sherkawi et al. presented their rule-based model to detect Arabic Speech Act types (Sherkawi et al., 2017). The Expert System has been developed in a bootstrapping manner, to classify an utterance written in the Modern Standard Arabic (MSA) to one of the sixteen speech act types (Affirmation, Negation, Confirmation, Interrogation, Imperative, Forbidding, Wishing, Vocative, Prompting, Rebuke, Exclamation, Hope, Condition, Praise, Dispraise, Swear). The system was tested on a hand-crafted corpus of about 1500 MSA sentences.

In a following research, (Sherkawi et al., 2018) proposed a statistical based technique to recognize MSA Arabic speech acts. The proposed technique used surface features, cue words and contextual information. The authors compared the results of multiple machine learning algorithms (Decision Trees, Naïve Bayes, Neural Networks and SVM) on a corpus of 1500 MSA sentences. The Decision Tree algorithm had the best results.

(Elmadany et al., 2018) used the JANA corpus (4725 utterances in Egyptian Dialect) to create a statistical dialogue analysis model for recognizing utterances dialogue acts using a machine learning approach based on multi-classes hierarchical structure.

In (Graja et al., 2013), authors used the TuDiCoI corpus (12182 utterances in Tunisian Dialect) to develop a discriminative algorithm based on conditional random fields (CRF) to semantically label spoken Tunisian dialect turns which are not segmented into utterances.

(Shala et al., 2010) applied speech act classification for Arabic discourse using SVM, NB and Decision Trees machine learning classifiers on a dataset of about 400 MSA utterances collected from newspapers.

One more work on Arabic language was conducted by (Hijjawi et al., 2013) whose approach was based on Arabic function words (such as, هل do, كيف how) . They focused on questions/non-questions utterance classification using decision trees.

To the best of our knowledge, there are no studies on the Dialogue Act recognition of Levantine Arabic Dialect.

3 Our Approach

Our system is built to be domain independent, but in this work, we have applied it on both restaurants order and airline ticketing systems. Hereafter, we will introduce our taxonomy, our in-house built datasets, preprocessing steps, and the different machine learning algorithms used.

3.1 Our Taxonomy

Based on our chatbot system, we have adopted our own taxonomy of speech acts that are mostly used in restaurants orders and airline ticketing. We divided the utterances into 8 types: (Greeting, Goodbye, Thanks, Confirm, Negate, Ask_repeat, Ask_for_alt, and Apology).

Table 1 presents the descriptions of our taxonomy with corresponding examples.

3.2 Our Dataset

To our knowledge, there is no available corpus in the case of Levantine dialect that can be used to develop our dialogue system. Therefore, we manually built our own dataset, which consists of sentences from two domains: Restaurants Orders and Airplane Ticketing domain.

Our corpus contains a set of 873 sentences that were manually tagged. We started from scratch and collected the sentences from different sources:

- (63%) Obtained by means of crowdsourcing: We asked our colleagues to write sentences of how they would imagine a restaurant order or flight reservation conversation would go, then we manually tagged the sentences according to our taxonomy.
- (32%) Extracted from Levantine tweets related to the two domains: A python code was used to download tweets according to keywords for every class, these sentences were then manually labeled.
- (5%) A dataset collected in a previous food order chatbot project (Shbib et al., 2017).

Dialogue Act	%	Description	Example Utterance
Greeting	12.9	Greeting a person and saying hello.	مرحبا كيفك شو أخبارك؟ marHaba kyfak $w0 >xbarak Hello how are you what are you up to?
Goodbye	11.0	Ending a conversation or saying goodbye.	وعليكم السلام الله معك wA Ealaykum alsalam Al~A maEak Peace be upon you, goodbye
Thanks	13.0	Thanking a person.	شكراً كتير $ukran ktyr Thanks a lot.
Confirm	13.6	Confirming a yes/no question.	أي أكيد >y >akyd Yes of course
Negate	11.8	Negating a yes/no question.	لا ما بدي lA mA bid~i No I don't want it
Ask_repeat	12.7	Asking the speaker to repeat what he said.	ممكن تعيد شو قلت؟ Mumkin tEyd $w qlt Can you repeat what you said?
Ask_for_alt	12.6	Asking for alternative options if given.	شوفي عندك غير خيارات؟ $w fi Eandak gyr xayarat What other options do you have?
Apology	12.0	Apologizing to a person.	آسف lsif Sorry

Table 1: Our Dialogue Acts Taxonomy.

In another experiment, we have created a multi-labeled version of the dataset in order to apply multi-labeling classification techniques to the task. The dataset has been manually retagged such that each sentence can belong to one or more class (Dialogue act). For both experiments, the data was divided 80% for training and 20% for testing.

3.3 Preprocessing

Different steps were taken to preprocess the data. First, data was resampled to create equal number of sentences for each class.

No stop words were removed because stop words like (yes/ نعم , no/لا, Ok/ماشي...) are key features in the classification of speech acts.

We also tested the impact of using the stem of the words vs. the full form words, on the Dialogue acts classification. Therefore, we used the Arabic ISRI stemmer and compared the results using SVM classification algorithm.

3.4 Classification Algorithms

We used a set of different classifiers with different features and compared them. The classification algorithms that were tried were LogisticRegression (LR), Support Vector Machine (SVM), MultinomialNB (MNB), ExtraTreesClassifier (ET), and RandomForestClassifier (RF). We also used the voting ensemble method to make the best prediction from each classifier.

The features that were tried in this paper are TF-IDF, N-gram (N-grams were tried from 1 to 5), a combination of TF-IDF and N-gram. We also compared some feature selection methods such as Select From Model, Feature Union, and Recursive Feature Elimination (RFE). We implemented different experiments, and assessed their results using precision, recall and f-measure metrics.

The comparison results of the N-gram feature on Logistic Regression classifier is shown in Table 2. The table shows that 2-gram is the best feature with an accuracy of 0.89%.

Ngram	1	2	3	4	5
Accuracy	0.88	0.89	0.87	0.87	0.87

Table 2: Accuracy using Logistic Regression with N-gram (1-5)

In order to minimize the number of features in our model, and only select the best features, we compared some feature selection models and tested their results on our Logistic Regression classifier. Results are shown in Table 3.

Feature Selection Model	LR Accuracy
Select From Model (Extra Trees Classifier)	89%
Select From Model (Random Forest Classifier)	88%
Select From Model (Linear SVC)	88%
Select K-best (k = 800)	91%
Select Percentile (percentile = 50)	89%

Table 3: Results of different Feature Selection Models

The results presented in Table 3 show that Select k-best with k = 800 feature is the best feature selection model, thus it will be used in our next experiments.

4 Evaluation

In order to evaluate our approach, we implemented five machine learning models: LogisticRegression (LR), Support Vector Machine (SVM), MultinomialNB (MNB), ExtraTreesClassifier (ET), and RandomForestClassifier (RF).

We trained each classifier on different features and compared the results. The voting ensemble method was also evaluated for each feature. Table 4 compares the results obtained using our models.

	N-gram	TF-IDF	TF-IDF & N-gram
LR	**0.91**	0.89	0.86
SVM	0.89	0.86	0.85
RF	0.79	0.73	0.72
MNB	0.86	0.87	0.85
ET	0.88	0.87	0.86
Voting	0.90	0.89	0.87

Table 4: Results of different Machine Learning Models

The results show that Logistic Regression model using N-gram features outperforms the rest. Logistic Regression model improved the Dialogue Acts labeling accuracy over the SVM model by 2%.

To study the impact of using a stemmer in the preprocessing step, we used the ISRI stemming algorithm which is implemented for Modern Standard Arabic, and to our knowledge there is no stemmer for the Levantine Dialect. Results showed that using the MSA stemming did not improve the accuracy of the recognition. The MSA stemmer produces incorrect stems such as ابد/Abad for the word مابدي/mAbid y, and لمو/lamw for the word يسلمو/yislamw. These erroneous stems will be part of the features used, and will definitely affect the classification results.

In order to further analyze the results, we looked into the confusion matrix to know which labels are correctly/incorrectly assigned to sentences.

Figure 1 shows the confusion matrix of our Logistic Regression. We notice that the most errors were made in sentences that belong to the class "Thanks" and were predicted as "confirm".

	Confirm	Ask_Repeat	Goodbye	Apology	Negate	Ask_For_Alt	Greeting	Thanks
Confirm	23	0	0	0	**2**	**1**	0	**2**
Ask_Repeat	**2**	30	0	**2**	**3**	0	0	0
Goodbye	**1**	0	26	0	**2**	0	0	0
Apology	**2**	0	0	34	**1**	0	0	0
Negate	**1**	**1**	0	0	34	0	0	**1**
Ask_For_Alt	**1**	0	0	0	**1**	44	0	0
Greeting	**1**	0	1	0	0	0	38	0
Thanks	**4**	0	0	**1**	0	**1**	0	30

Figure 1: Confusion Matrix of the LR model

We noticed from the false predicted utterances that the sentences in fact belong to both classes, "Thanks" class and "Confirm" class. Table 5 shows some examples of these mislabeled sentences.

Dialog Act	Sentence
Thanks	أي ماشي يعطيك العافية بس لا تتأخرو بالتوصيل
Thanks	أي شكرا يعطيك العافية بس التوصيل اديش بيكلف
Thanks	أي يعطيك العافية

Table 5: Examples of "Thanks" sentences predicted as "Confirm"

To solve this problem, we re-labeled the data so each sentence would belong to more than one class, then we applied the One Vs. Rest multi-labeling classifier. The Results using different classifiers are shown in Table 6

Results show that our SVM classifier outperforms the rest of the classifiers with an accuracy of 86%.

One Vs. Rest Classifier	Accuracy
LR	0.84
SVM	**0.86**
RF	0.84
MNB	0.85
ET	0.82

Table 6: Results of the different Multi-labeling classifiers

5 Conclusion

In this paper, we have investigated different Dialogue act recognition models for Levantine Arabic language. The best model will be embedded into the Language Understanding component in our Arabic Conversational (Syrian Levantine Dialect) system.

We implemented different Machine Learning algorithms along with different features and feature selection methods. We evaluated the proposed techniques on a hand-crafted dataset in the restaurant's orders and airline ticketing domain. The best results were achieved using SVM model with 86% accuracy).

In the future, we intend to record a real restaurant and Ticket ordering conversations and create a new larger dataset, with real life situations and speech act sequences. This new dataset will allow us to take into consideration the whole context of the sentence in predicting the speech act of each utterance.

Building a Morphological Analyzer (or even a simple light stemmer) for Levantine (Syrian) Arabic, and using it in the preprocessing steps, will allow to extract many important features such as dialect negation tools (usually concatenated with the word itself, such as مارح/I will not, مابدي/I don't want), and this will improve the correct dialogue acts recognition.

References

John Langshaw Austin. 1962. *How to do things with words*. Harvard University Press, Cambridge.

Chandrakant Bothe, Cornelius Weber, Sven Magg, and Stefan Wermter. 2018. A context-based approach for dialogue act recognition using simple recurrent neural networks. Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Christophe Cerisara, Pavel Král, and Ladislav Lenc. 2018. On the effects of using word2vec representations in neural networks for dialogue act recognition.

Zheqian Chen, Rongqin Yang, Zhou Zhao Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. pages 225–234. The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval ACM.

AbdelRahim Elmadany, Sherif Abdou, and Mervat Gheith. 2018. Improving dialogue act classification for spontaneous arabic speech and instant messages at utterance level. In *The 11th edition of the Language Resources and Evaluation Conference.*

Aysu Ezen-Can and Kristy Elizabeth Boyer. 2014. Combining task and dialogue streams in unsupervised dialogue act models. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 113–122.

Mark Fišel. 2007. Machine learning techniques in dialogue act recognition. pages 117–134.

Toshiaki Fukada, Detlef Koll, Alex Waibel, and Kouichi Tanigaki. 1998. Probabilistic dialogue act extraction for concept based multilingual translation systems. In *The Fifth International Conference on Spoken Language Processing*.

Marwa Graja, Maher Jaoua, and Lamia Hadrich Belguith. 2013. Discriminative framework for spoken tunisian dialect understanding. In *The International Conference on Statistical Language and Speech Processing*, pages 102–110. Springer.

Nizar Habash. 2010. *Introduction to Arabic natural language processing*. Synthesis Lectures on Human Language Technologies.

Mohammad Hijjawi, Zuhair Bandar, and Keeley Crockett. 2013. User's utterance classification using machine learning for arabic conversational agents. In *The 5th International Conference on Computer Science and Information Technology*, pages 223–232. IEEE.

Zoleikha Jahanbakhsh-Nagadeh, Mohammad-Reza Feizi-Derakhshi, and Arash Sharifi. 2019. A speech act classifier for persian texts and its application in identify speech act of rumors.

Hamed Khanpour, Nishitha Guntakandla, and Rodney D. Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Khanpour, Hamed, Nishitha Guntakandla, Rodney D. Nielsen*. COLING.

Hark-Soo Kim, Choong-Nyoung Seon, and Jung-Yun Seo. 2011. Review of korean speech act classification: machine learning methods. 5:288–293.

Minkyoung Kim and Harksoo Kim. 2018. Dialogue act classification model based on deep neural networks for a natural language interface to databases in korean. In *IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 537–540. IEEE.

Aysu Ezen-Can Kristy Elizabeth Boyer. 2015. A tutorial dialogue system for real-time evaluation of unsupervised dialogue act classifiers: Exploring system outcomes. In *Proceedings of the international conference on artificial intelligence in education (AIED), Lecture Notes in Computer Science*, pages 105–114. Springer.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, Sachindra Joshi, and Arun Kumar. 2018. Dialogue act sequence labeling using hierarchical encoder with crf. The Thirty-Second AAAI Conference on Artificial Intelligence.

Lee, Ji Young, and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In *NAACL*. arXiv preprint arXiv:1603.03827.

Shereen Oraby, Pritam Gundecha, Jalal Mahmud, Mansurul Bhuiyan, and Rama Akkiraju. 2017. How may i help you?: Modeling twitter customer serviceconversations using fine-grained dialogue acts. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 343–355. ACM.

Vasile Rus, Nabin Maharjan, and Rajendra Banjade. 2017. Dialogue act classification in human-to-human tutorial dialogues. In *Innovations in Smart Learning*, pages 185–188, Singapore. Springer.

Lubna Shala, Vasile Rus, and Arthur C. Graesser. 2010. Automatic speech act classification in arabic. In *Subjetividad y Procesos Cognitivos Conference*, pages 284–292.

Boushra Shbib, Batool Ibo, Dima Qawoq, and Safa Alshaib. 2017. Arabic conersational agent for foor ordering.

Lina Sherkawi, Nada Ghneim, and Oumayma Al Dakkak. 2017. Arabic speech act recognition using bootstrapped rule based system.

Lina Sherkawi, Nada Ghneim, and Oumayma Al Dakkak. 2018. Arabic speech act recognition techniques.

Maryam Tavafi, Yashar Mehda, Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2013. Dialogue act recognition in synchronous and asynchronous conversations. In *Proceedings of the SIGDIAL 2013 Conference*, page 117–121.

Wan, Yao; Yan, Wenqiang, Jianwei, Gao; Zhao, Zhou ; Wu, Jian; S. Yu, and Philip. 2018. Improved dynamic memory network for dialogue act classification with adversarial training. In *The 2018 IEEE International Conference on Big Data (Big Data)*, pages 841–850. IEEE.

Elina Zarisheva and Tatjana Scheffler. 2015. Dialog act annotation for twitter conversations. In *Proceedings of the SIGDIAL 2015 Conference*, page 114–123.